

Kedma Batista Duarte

ASSESSING RESEARCHER QUALITY FOR  
COLLABORATIVE PURPOSES

Thesis submitted to the Graduate Program in Engineering and Knowledge Management at the Universidade Federal de Santa Catarina, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Engineering and Knowledge Management.

Advisor: Prof. Roberto C.S. Pacheco, PhD

Co-advisor: Prof<sup>a</sup>. Rosina O. Weber, PhD

Florianopolis  
2017

Ficha de identificação da obra elaborada pelo autor,  
através do Programa de Geração Automática da Biblioteca Universitária da UFSC

Duarte, Kedma Batista

Assessing researcher quality for collaborative  
purposes / Kedma Batista Duarte ; orientador,  
Roberto Carlos dos Santos Pacheco, coorientadora,  
Rosina O Weber, 2017.

247 p.

Tese (doutorado) - Universidade Federal de Santa  
Catarina, Centro Tecnológico, Programa de Pós  
Graduação em Engenharia e Gestão do Conhecimento,  
Florianópolis, 2017.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2.  
Career trajectories. 3. Case-based Reasoning. 4.  
Knowledge Engineering. 5. Research collaborator .  
I. Pacheco, Roberto Carlos dos Santos. II. Weber,  
Rosina O. III. Universidade Federal de Santa  
Catarina. Programa de Pós-Graduação em Engenharia e  
Gestão do Conhecimento. IV. Título.

Kedma Batista Duarte

ASSESSING RESEARCHER QUALITY FOR  
COLLABORATIVE PURPOSES

This Doctoral Thesis was considered adequate for granting its author the Title of “Ph.D. in Engineering and Knowledge Management”; its final version was approved by the Graduate Program in Engineering and Knowledge Management at the Universidade Federal de Santa Catarina.

Florianópolis, 22<sup>nd</sup> November 2017.



---

Prof. Gertrudes Aparecida Dandolini, PhD  
Program Coordinator

Examining Committee:



---

Prof. Roberto Carlos dos Santos Pacheco, PhD  
Advisor  
Universidade Federal de Santa Catarina, Brasil



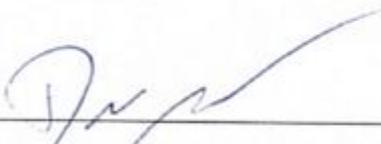
---

Prof. Rosina O. Weber, PhD  
External Co-Advisor  
Drexel University, USA



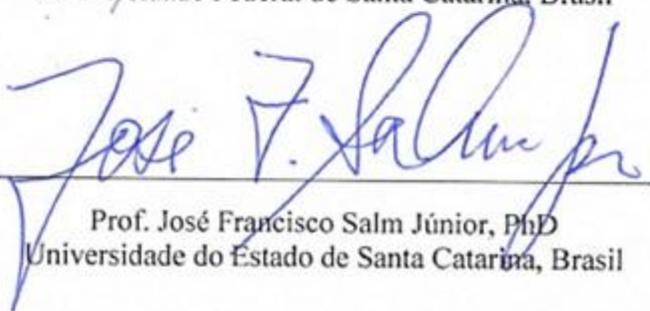
---

Prof. Alexandre Leopoldo Gonçalves, PhD  
Universidade Federal de Santa Catarina, Brasil



---

Prof. Denilson Sell, PhD  
Universidade Federal de Santa Catarina, Brasil



---

Prof. José Francisco Salm Júnior, PhD  
Universidade do Estado de Santa Catarina, Brasil



---

Prof. Silvestre Labiak Jr, PhD  
Universidade Tecnológica Federal do Paraná, Brasil



---

Prof. Marcelo Albano Moret Simões Gonçalves, PhD  
Universidade do Estado da Bahia, Brasil

I dedicate this thesis to

My father, Antônio Duarte and my brother, Julio Carlos Duarte, who always encouraged me to pursue my dreams. They are and always will be my source of inspirations, my references of values and principles of life, my eternal heroes (*in memoriam*).

My mom, Dalva Duarte, who first taught me about collaboration. I affectionately thanks for strengthen my convictions, and my determination to study this fascinating theme.



Eu dedico essa tese ao

Meu pai, Antônio Duarte e meu irmão, Julio Carlos Duarte, que sempre me encorajaram a perseguir meus sonhos. Eles são e sempre serão minha fonte de inspiração, minhas referências de valores e princípios da vida, meus eternos heróis (*in memoriam*).

Minha mãe, Dalva Duarte, que primeiro me ensinou sobre colaboração. Agradeço carinhosamente por fortalecer minhas convicções e minha determinação em estudar esse fascinante tema.



## ACKNOWLEDGEMENTS

Above all, thanks God for the grace to live these amazing PhD years. I am grateful for sustain me during the course of this thesis.

In special, I thank Professor Roberto Pacheco, my principal advisor, for receive me at the Engineering and Knowledge Management graduate program. Thanks for present me the transdisciplinary science with his visionary mind, and show me new paths in my career trajectory. I also would like to thank for his valuable support, advice, and friendship, which were essential for my choices.

My sincere gratitude to my co-advisor, Professor Rosina Weber, who encouraged me to investigate the research collaborators. Thanks for make me think deeper, and for her patience in teach me about science. Thanks for motivate me to go beyond my own borders. A special thank for her friendship, by receiving and supporting me in Philadelphia, and at Drexel University.

I kindly thank Professor Paulo Selig, for introduce me the Engineering and Knowledge Management graduate program, in the year of 2009. Thanks for collaborate with the happiest years of my life.

In particular, I would like to express my appreciation for Professor Maria Zaira Turchi, for trusting me, and allowing me to live this challenge. Thanks for help me along my PhD, supporting and contributing with my dream of becoming a researcher.

Thanks also go to my precious friend Professor José Clecildo Barreto Bezerra, who introduced me this fascinating world of Science, Technology and Innovation. I am grateful for teach me about research collaboration, and show me the wealth of promoting the collaboration among researchers.

Thanks go to Victor Santibanez, my English teacher, my great friend, who lead me to write this thesis in English, and not only this, but provided me in each step of the thesis, valuable suggestions. Thanks for your feedback in each meeting.

Then, I thank my family, my sister Sheila, Rhicardo, Felipe and Fabíola, Lucas and Kelly, and Isabela, Lineker and the baby. Their enthusiasm and emotional support incentivized me along these five years. I extend my thanks to Leandro, Giselle, Matheus and Lara, my cousins in Florianópolis.

I give thanks for had meet so careful friends Neusa and Orlando Coutinho, who introduced me in so special Presbyterian community. You are my family in Florianópolis. God bless you!

Thank my precious friends Margareth and Arioberto, Wilda and Wanderson, Jaiza Alves, Angela Marisa, Marcia Schiavon, and Renata Speltri, who visited me in Florianópolis.

In special, thanks Professor Jane Greenberg, who gave me the opportunity to present my work to the Metadata Research Center (MRC) at College of Computing and Informatics, Drexel University, which I visited during my PhD.

So, I give also special thanks those friends that received me in Philadelphia, in special Jonathan Newmark and Rosina; Matt Miller; Leonardo and Paula; Rodolfo and Larissa.

Thanks Stela Institute, in special to Professors Aran Morales, José Leomar Todesco, Denilson Sell, José Salm Jr, and Roberto Pacheco. In addition, Fernando Montenegro, and Marcos Marchezan. Thanks all, for support my studies during my PhD, allowing me to investigate the *Lattes* database. In particular, thanks Ruderger Taxweiler for the effort with the data extraction. Thanks, in especial go to Viviane, Sandra, and Alessandra, and all friends that made me felt part of this team.

My satisfaction to belong to the Engineering and Knowledge Management graduate program (EGC). Thanks go to all professors, graduate students, and staff, for these five years of knowledge sharing. In the EGC I had meet friends closer than brothers, and some of them are, Julieta Wilbert, Viviane Schneider, Vivian Alves, Micheline Krause, Paula Campos, Cinthya Zanuzzi, Silvia Bentancourt, and Julio Casaes. Each one of you is in some way inside my thesis. Silvia and Julio, I will never forget the planet Venus and the Planktons.

Many, many thanks to my friends from Goiânia, who despite the distance, followed my steps, encouraging and praying for me. Cheers to all my friends, who I am not mention here, but I look forward to the opportunity to apologize and say, “*thank you*” in person.

Finally, my recognition to Goiás State Research Support Foundation (FAPEG) and Goiás State University (UEG) for the financial support, under agreement number 201310267000099.

## AGRADECIMENTOS

Acima de tudo, agradeço a Deus pela graça de viver estes maravilhosos anos de doutorado. Sou grata por Ele me sustentar durante a condução desta tese.

Em especial, agradeço ao Professor Roberto Pacheco, meu principal orientador, por me receber no Programa de Pós-graduação em Engenharia e Gestão do Conhecimento. Obrigado por me apresentar à ciência transdisciplinar, com sua mente visionária, e mostrar me novos caminhos em minha trajetória de carreira. Eu também gostaria de agradecer pelo seu valioso apoio, conselhos e amizade, os quais foram essenciais para minhas escolhas.

Minha sincera gratidão à minha co-orientadora, Professora Rosina Weber, que me encorajou a investigar os colaboradores científicos. Obrigada por me fazer pensar com profundidade, e por sua paciência em me ensinar sobre a ciência. Agradeço por me motivar a ir além das minhas próprias fronteiras, e em especial por sua amizade, me recebendo e apoiando quando estive na Universidade de Drexel, na Filadélfia.

Meu carinhoso agradecimento ao Professor Paulo Selig, por ter a me apresentado o programa de pós-graduação em Engenharia e Gestão do Conhecimento, no ano de 2009. Obrigada por colaborar com os anos mais felizes da minha vida.

Em particular, gostaria de expressar o meu apreço pela Professora Maria Zaira Turchi, por sua confiança em mim, e por me permitir viver esse desafio. Obrigado por me ajudar ao longo do meu doutorado, apoiando e contribuindo com meu sonho de me tornar uma pesquisadora.

Agradeço também ao meu precioso amigo, o Professor José Clecildo Barreto Bezerra, que me apresentou a este fascinante mundo da ciência, tecnologia e inovação. Eu sou grata por me ensinar sobre colaboração científica, e me mostrar a riqueza de promover a colaboração entre os pesquisadores.

Agradeço ao Victor Santibanez, meu professor de inglês, meu grande amigo, que me levou a escrever esta tese em inglês, e não só isso, mas me forneceu sugestões valiosas em cada etapa da tese. Obrigada pelo seu feedback nas divertidas e agradáveis aulas.

Então, agradeço a minha família Sheila, Rhicardo, Felipe e Fabíola, Lucas e Kelly, Isabela, Lineker e o bebê que está nascendo, Djair e Maraísa. O entusiasmo e apoio emocional de vocês me incentivou ao

longo desses cinco anos. Eu estendo meus agradecimentos aos meus primos em Florianópolis, Leandro, Giselle, Matheus e Lara.

Eu dou graças por ter encontrado amigos tão cuidadosos, Neusa e Orlando Coutinho, que me apresentaram à uma comunidade presbiteriana tão especial. Vocês são minha família em Florianópolis. Que Deus vos abençoe!

Obrigada meus preciosos amigos goianos, Margareth e Arioberto, Wilda e Wanderson, Jaiza, Ângela Marisa, Márcia Schiavon, Renata Speltri, e Flávio Rodrigues, que me visitaram em Florianópolis.

Em especial, agradeço a Professora Jane Greenberg, pela oportunidade de apresentar meu trabalho no grupo de pesquisa, Metadata Research Center (MRC), College of Computing and Informatics, Drexel University, que visitei nos Estados Unidos, durante o meu doutorado.

Então, agradeço os amigos que me receberam na Filadélfia, em especial Jonathan Newmark e Rosina; Matt Miller; Leonardo e Paula; Rodolfo e Larissa.

Agradeço ao Instituto Stela, em especial aos Professores Aran Morales, José Leomar Todesco, Denilson Sell, e José Salm Jr. Além destes, ao Fernando Montenegro e Marcos Marchezan. Agradeço a todos, por apoiar meus estudos durante o período de doutorado, permitindo-me investigar a base de dados Lattes. Em particular, agradeço ao amigo Rudger Taxweiler pelo esforço com a extração de dados. Agradeço ainda à Sandra, Alessandra, e todos os amigos que me fizeram sentir parte desta equipe.

Tenho orgulho de pertencer ao Programa de Pós-graduação em Engenharia e Gestão do Conhecimento (EGC). Agradeço a todos os professores, colegas de pós-graduação e funcionários, por esses cinco anos de compartilhamento de conhecimento. No EGC eu conheci amigos mais próximos que irmãos, e alguns deles são, Julieta Wilbert, Viviane Schneider, Vivian Alves, Micheline Krause, Paula Campos, Cinthya Zanuzzi, Silvia Bentancourt, e Julio Casaes. Cada um de vocês está de alguma forma dentro da minha tese. Silvia e Julio nunca esquecerei o planeta Venus e os Planktons.

Muitos, muitos agradecimentos aos meus amigos de Goiânia, que apesar da distância, seguiram meus passos, me encorajando e orando por mim. Agradecimentos a todos os amigos, que não menciono aqui, mas que aguardo com expectativa a oportunidade de me desculpar e dizer um carinhoso "*obrigado*" pessoalmente.

Finalmente, meu reconhecimento à Fundação de Amparo à Pesquisa de Goiás (FAPEG) e à Universidade Estadual de Goiás (UEG) pelo apoio financeiro, sob o número de processo 201310267000099.

**Salmo 139,16**

**sob  
teus olhos eu in-  
forme, inscrito  
qdo nem uma  
letra sobre-te  
aqui  
s via.**

**(Poemas de Júlio Carlos Duarte, 1991)**



## RESUMO

Avaliar a qualidade do pesquisador tem sido um desafio constante para os tomadores de decisão, os quais precisam de métodos mais eficientes, baseados em critérios objetivos, para orientar políticas de pesquisa. Por exemplo, em propósitos tais como recrutamento, promoção e fomento. Esta tese investiga a avaliação da qualidade do pesquisador, tendo como principal foco, os colaboradores científicos. Atualmente, as métricas para avaliar a colaboração científica são baseadas no *índice de citações* e em *coautoria*. No entanto, a literatura tem recomendado investigar métodos para mensurar a colaboração científica que vão além das taxas de citação. Outro fato é que, no processo de seleção de pesquisadores individuais para fins colaborativos, considerar todo o grupo não é suficiente, uma vez que os indivíduos devem ser avaliados e não o grupo inteiro, como no caso das redes de coautoria. Além disso, críticas sobre a aplicação incorreta de métricas tem incitado um debate entre pesquisadores e tomadores de decisão para o uso correto de métricas de pesquisa. Assim, esta tese propõe, um método que considera indivíduos e o propósito da avaliação, o *purpose-oriented method*. Esta solução baseia-se na Engenharia do Conhecimento, adota Raciocínio Baseado em Casos como metodologia de implementação, e usa dados da base de dados *Lattes*. O método proposto avalia automaticamente a qualidade dos pesquisadores, aplicando medidas de similaridade aos seus *curriculum vitae*, considerando a experiência de pesquisadores bem-sucedidos para avaliar pesquisadores candidatos a um processo de seleção alvo. Os resultados de dois cenários experimentais demonstram a usabilidade do método proposto, bem como, as contribuições desta tese para os tomadores de decisão em Ciência e Tecnologia. O estudo contribui com uma metodologia que demonstra “*como fazer*” para mensurar a qualidade dos colaboradores científicos com base em suas trajetórias de carreiras. Além disso, a solução permite comparar automaticamente um grande número de *curriculum vitae*, apoiando avaliações de especialistas qualitativos. Acima de tudo, este estudo contribui com pesquisadores e tomadores de decisão, aumentando a compreensão da avaliação individual de pesquisadores em propósitos colaborativos.

**Palavras-chave:** Trajetórias de carreira. Raciocínio Baseado em Casos. Engenharia do Conhecimento. Colaboração Científica. Colaborador Científico.



## RESUMO EXPANDIDO

### Introdução

"Uma boa ciência só pode acontecer com bons cientistas". Apesar deste evidente requisito, declarado por Collins, Morgan e Patrinos (2003), determinar a qualidade do pesquisador é um grande desafio para os tomadores de decisão na ciência (VAN NOORDEN et al., 2013). Assim, é crescente a procura por métodos eficientes de mensuração para propósitos tais como, recrutamento, promoção e decisões de fomento (HAUSTEIN; LARIVIÈRE, 2015; LANE, 2010).

O desafio de mensurar a qualidade do pesquisador por critérios mais objetivos, foi intensificado na década de 1960 (OKUBO, 1997). Este foi um período de incentivos ao avanço científico e tecnológico, como por exemplo, a criação das agências de fomento (NARIN; HAMILTON, 1996; OKUBO, 1997). No entanto, foi também um período em que tomadores de decisão perceberam que a ciência precisava de mais recursos para enfrentar as necessidades da humanidade e, portanto, seriam necessários critérios mais objetivos para orientar a tomada de decisão em políticas de pesquisa científica (NARIN; HAMILTON, 1996).

O índice de citações, *Science Citation Index (SCI)*, foi o primeiro indicador bibliométrico desenvolvido para mensurar a qualidade do pesquisador (GARFIELD, 1964). O conceito por trás do SCI é que o número de artigos publicados por um pesquisador, oferece de fato alguma medida de sua atividade (GARFIELD; MALIN, 1968). O SCI é também objeto de uma vasta literatura, a qual inclui a mensuração da qualidade do colaborador científico, tais como, coautoria de artigos científicos (*coauthorship*) (BEAVER; ROSEN, 1978), e rede de coautorias (*coauthorship networks*) (NEWMAN, 2001, 2004).

Métricas objetivas são cada vez mais utilizadas para quantificar a qualidade científica (VAN NOORDEN, 2010), no entanto, apesar das vantagens de sua adoção, críticas sobre sua incorreta aplicação tem levado os pesquisadores a um debate. Por exemplo, indicadores projetados para avaliar periódicos têm sido erroneamente aplicados na avaliação de indivíduos e grupos (DAVID; FRANGOPOL, 2015).

Em resposta a tais críticas, um grupo de pesquisadores propôs um conjunto de princípios denominado *Leiden Manifesto for research metrics* (HICKS et al., 2015). Tais princípios recomendam o uso de métricas objetivas para o apoio a julgamentos qualitativos, e sugerem que tais métricas sejam alinhadas à missão e objetivos das instituições. Eles enfatizam a necessidade de reconhecimento das pesquisas locais,

declaram que as métricas devem ser transparentes, e ressaltam a importância da qualidade dos dados e de sua atualização. Além disso, apontam a necessidade de se rever o portfólio dos pesquisadores e alertam sobre os riscos de resultados injustos quando usadas erroneamente.

Em relação à colaboração científica, Bozeman, Fay e Slade (2013) revê a literatura e recomenda que deveria ser dada mais atenção à falta de métodos para mensurar a colaboração científica. Para estes autores a pesquisa sobre colaboração científica deveria encontrar uma melhor forma de mensurar o impacto do conhecimento, indo além das taxas de citação (BOZEMAN; FAY; SLADE, 2013).

Esta tese, investiga o problema da mensuração da qualidade do pesquisador, com atenção especial aos colaboradores científicos, em processos de seleção de pesquisadores, tais como, decisões de recrutamento, promoção e fomento.

Assim, considerando as necessidades de instrumentos de governança para guiar decisões de políticas científicas, os princípios declarados no Leiden Manifesto (HICKS et al., 2015), a recomendação de Bozeman, Fay e Slade (2013) para colaboração científica, esta tese tem como objetivo responder a seguinte questão de pesquisa "*Como mensurar a qualidade do pesquisador para propósitos colaborativos?*". Esta questão de pesquisa é apoiada em duas sub questões: "*RQ1: Como conceitualizar um modelo de dados para mensurar a qualidade do pesquisador com ênfase em colaboradores científicos?*", e "*RQ2: Como avaliar a qualidade do pesquisador?*".

## **Objetivos**

Considerando as questões de pesquisa anteriormente mencionadas, esta tese tem como objetivo geral propor um método para mensurar a qualidade do pesquisador em propósitos colaborativos. Desta forma, visando orientar a condução deste estudo, os seguintes objetivos específicos são propostos:

1. Identificar fatores, conceitos e elementos adequados à avaliação de pesquisa em colaboração;
2. Identificar métodos e técnicas de Engenharia do Conhecimento, que podem ser utilizados para implementar soluções orientadas para fins de avaliação da qualidade de pesquisador;
3. Desenvolver um método capaz de contribuir com a avaliação da qualidade do pesquisador, particularmente em fins colaborativos.

## Metodologia

Esta tese baseia-se em uma visão de mundo quantitativa, na qual, variáveis podem ser medidas por instrumentos, e dados podem ser analisados por estatística (CRESWELL, 2009). O estudo se posiciona como uma pesquisa aplicada, que busca objetivos específicos, e resultados aplicados em problemas práticos (OCDE, 2015b). Além disso, a tese concentra-se em abordagens focadas a colaborações interdisciplinares, caracterizadas pela colaboração entre pesquisadores de diferentes disciplinas, com uma metodologia comum e um problema compartilhado (MOBJÖRK 2010).

O procedimento metodológico é conduzido em cinco etapas: Primeiro o problema é definido, e objetivos específicos formulados. Em seguida, a literatura é revista a fim de examinar quatro constructos: *Mensuração de pesquisa*; *Colaboração científica*; *Fontes de conhecimento*; e *métodos e técnicas da Engenharia de Conhecimento*. Depois, o domínio de conhecimento sobre pesquisadores e colaboradores científicos é investigado, com o resultado sendo representado por um modelo conceitual de dados. Então, a solução proposta é projetada e implementada, a qual é chamada *Purpose-oriented method*. Esta etapa inclui a especificação dos dados utilizados neste estudo. Por fim, o método proposto é introduzido por meio de dois experimentos, os quais demonstram sua usabilidade em cenários de ciência e tecnologia (C&T).

O escopo do estudo foca na *mensuração da pesquisa*, particularmente na investigação dos *colaboradores científicos*. O estudo adota o conceito clássico definido por Katz e Martin (1997), no qual *colaboração científica* é entendida como o trabalho conjunto de pesquisadores para alcançar o objetivo comum de produzir novos conhecimentos científicos. Esta tese também considera que *qualidade* é “*fitness for purpose*” (JURAN; GODFREY, 1999), conceituando qualidade como dependente de perspectivas, necessidades e prioridades de usuários. Além disso, os 10 princípios do Leiden Manifesto (HICKS et al., 2015) são levados em conta como fontes de referência.

Métodos de *Engenharia do Conhecimento (EC)*, tais como *Raciocínio Baseado em Casos (RBC)* são adotados no desenvolvimento do método proposto, e a base de dados *Lattes* ([lattes.cnpq.br](http://lattes.cnpq.br)) é usada como fonte de currículos de pesquisadores, fornecendo dados de alta qualidade ao estudo (LANE, 2010). Por fim, o estudo adere ao Programa de Pós-graduação em Engenharia e Gestão do Conhecimento (PPGEGC/UFSC) como uma solução de EC aplicada à Gestão do Conhecimento, em organizações de C&T.

## Resultados e Discussão

*Purpose-oriented method*, a solução proposta que responde a questão de pesquisa, “*Como avaliar a qualidade do pesquisador para propósitos colaborativos?*”, é apresentado nesta tese. Esta solução implementa uma inovadora abordagem de método de avaliação, ou mais apropriadamente, de mensuração da colaboração científica, focando no colaborador científico.

O método proposto tem como cerne, um classificador o qual é chamado de *purpose-oriented classifier*, o qual é uma implementação de dois estágios do ciclo de RBC, *Retrieve* e *Reuse*, descrita em quatro estágios. (i) Descrição do problema; (ii) Aprendizado dos pesos para representar o propósito da avaliação; (iii) Classificação dos pesquisadores candidatos como aptos (*fit*) ou inaptos (*Unfit*) ao propósito da avaliação; (iv) Ranqueamento dos pesquisadores candidatos, aptos ao propósito da avaliação. A execução destes quatro estágios é como *Purpose-oriented method* avalia a qualidade do pesquisador, cujos primeiros resultados realizados por esta tese foram obtidos por meio de dois experimentos.

Experimento I considerou a hipótese que “*Um método orientado a propósitos é mais acurado que um método de propósito independente*”. Este experimento mostrou o alinhamento da solução com o segundo princípio do Leiden Manifesto (HICKS et al, 2015), o qual argumenta que aspectos contextuais deveriam ser considerados em processos de avaliação. Para testar este experimento, três classificadores foram criados, um de propósito independente (PIC) e os outros dois orientados a propósito, POC1 orientado a colaboração, e POC2 orientado a trabalhos individuais. Ao final do experimento, o classificador de propósito independente classificou corretamente 62% dos candidatos para trabalhos em colaboração e 71% para trabalhos individuais. Já o classificador orientado a propósito classificou corretamente 92% e 93%, respectivamente, confirmando a hipótese.

Experimento II simulou em detalhes, um processo de recrutamento de pesquisadores candidatos para trabalhar em colaboração com membro de um grupo de pesquisa Brasileiro de reconhecida excelência em colaboração científica. Este experimento foi inspirado no sétimo princípio do Leiden Manifesto (HICKS et al, 2015). Ele incorporou o tratamento da trajetória de carreira dos pesquisadores no *purpose-oriented method*, tornando o método mais rigoroso. Por exemplo, na primeira fase do experimento, de 15266 candidatos, 3450 candidatos foram selecionados. Após considerar a trajetória de carreira do pesquisador, e levar em consideração os últimos cinco anos consecutivos

como aptos ao processo de avaliação, 1918 candidatos foram selecionados. Ao final, estes resultados confirmaram o pressuposto que “*incorporar o tratamento de trajetórias de carreiras no ‘purpose-oriented method’ leva à resultados melhor alinhados com os objetivos da avaliação*”.

## **Considerações Finais**

Os resultados dos experimentos, citados acima, demonstram a usabilidade e relevância do método proposto para a apoio às decisões em C&T, tais como recrutamento, promoção e fomento. Acredito que o estudo apresentado leva a uma grande quantidade de contribuições, porém irei focar a atenção em três que considero particularmente significantes.

A primeira contribuição aborda a entrega aos gestores de C&T, um novo método para avaliar a qualidade de pesquisadores. A solução resultante vai além da concepção inicial de um método orientado a propósitos, mas apresenta uma metodologia que descreve como avaliar colaboradores científicos, com base em suas trajetórias de carreira.

A segunda contribuição diz respeito à aplicabilidade da solução proposta em apoio aos processos de avaliações qualitativas, uma vez que eliminaria um grande volume de trabalho realizado pelos avaliadores. Por exemplo, *purpose-oriented method* compara automaticamente um grande número de currículos de pesquisadores, analisando cada ano de suas trajetórias de carreira. Nesta análise, são aplicadas medidas de similaridade, que contrastam pesquisadores candidatos com pesquisadores bem-sucedidos.

A terceira, e uma das maiores contribuições da tese, é a capacidade da solução representar os critérios da avaliação, por meio da relativa relevância dos atributos dos pesquisadores bem-sucedidos. Assim, os decisores obtém uma melhor compreensão do propósito colaborativo da avaliação. Por exemplo, a identificação e análise diferentes perfis de pesquisadores bem-sucedidos, bem como, de pesquisadores individuais em objetivos colaborativos.

**Palavras-chave:** Trajetórias de carreira. Raciocínio Baseado em Casos. Engenharia do Conhecimento. Colaboração Científica. Colaborador Científico.



## ABSTRACT

Assessing researcher quality has been a constant challenge for decision makers, who need more efficient methods, based on objective criteria, to guide research policy. For instance, in purposes such as, recruitment, promotion, and funding decisions. This thesis investigates researcher quality assessment, having as its main focus, the research collaborators. Currently, metrics to assess research collaboration are based on the *citation index*, and *co-authorship*. However, the literature has recommended investigating methods to measure research collaboration that go beyond the citation rates. Other fact is that, in selection processes of individual researchers for collaborative purposes, considering the entire group as a whole it is not enough, since individuals must be assessed and not the entire group. Moreover, criticisms regarding the misapplication of metrics have incited a debate between researchers and decision makers for the correct use of research metrics. Thus, this thesis proposes, an approach that considers individuals and the purpose of the assessment, *the purpose-oriented method*. This approach is based on Knowledge Engineering, by adopting Case-Based Reasoning methodology, and data from the Brazilian *Lattes* database. The *purpose-oriented method* automatically assesses researcher quality, by applying similarity measures to their *curriculum vitae*, considering the experience of successful researchers to assess candidate researchers to a target selection process. The results of two experimental scenarios, demonstrates the usefulness of the *purpose-oriented method*, as well as, the contributions of this thesis for decision makers in Science and Technology. The study contributes with a methodology that demonstrates “*how to do*” to assess research collaborators based on their career trajectories. Furthermore, the approach allows automatically to compare large numbers of researchers’ curriculum vitae, supporting qualitative expert assessments. Above all, this study contributes with researchers and decision makers by enhancing the comprehension of individual researcher assessment in collaborative purposes.

**Keywords:** Career trajectories. Case-based Reasoning. Knowledge Engineering. Research collaboration. Research collaborator.



## LIST OF FIGURES

Figure 1 – Methodological procedure .....	44
Figure 2 – Delimitations of this study .....	45
Figure 3 – The interdisciplinary adherence of this thesis to the PPGEGC/UFSC.....	47
Figure 4 – The structure of this document.....	49
Figure 5 – The four literature review constructs for this thesis .....	51
Figure 6 – The 10 principles of the Leiden Manifesto for research metrics .....	63
Figure 7 – The overview of the future of STI.....	64
Figure 8 – The study of dynamics of science from 1962 to 1980, according to Kuhn (1962).....	72
Figure 9 – An example of a small co-authorship network.....	78
Figure 10 – The essential components of a KBS.....	95
Figure 11 – Learning system basic structure .....	95
Figure 12 – The Methodological Pyramid.....	96
Figure 13 – The CommonKADS models. ....	97
Figure 14 – The groups and types of knowledge intensive tasks. ....	98
Figure 15 – The CBR cycle .....	100
Figure 16 – The CBR cycle combined with Machine learning techniques .....	102
Figure 17 – The Spy technique in S-EM .....	111
Figure 18 – The defuzzification process. ....	114
Figure 19 – Methodological procedure of this thesis .....	121
Figure 20 – The organizational environment in which the proposed method will operate .....	123
Figure 21 – The knowledge model of the context layer for this thesis	124
Figure 22 – The design model of the artifact layer for this thesis .....	125
Figure 23 – Growth of scientific collaboration from 1976-2014. ....	131
Figure 24 – The research collaborator conceptual data model includes the context where a research collaborator delivers scientific activities to achieve accomplishments.....	135
Figure 25 – The elements of the data description process.....	147
Figure 26 – Evolution of the Brazilian researcher accomplishments from 1950- 2015.....	155
Figure 27 – The <i>purpose-oriented method</i> to assess researcher quality .....	160
Figure 28 – The <i>purpose-oriented classifier</i> .....	162
Figure 29 – Stage 1: Describing the target problem of the research assessment .....	163

Figure 30 – Stage 2: Learning purpose through weights.....	164
Figure 31 – The problem of finding genuine successful CVs. ....	166
Figure 32 – The datasets used to the balanced set (M) initialization process.....	167
Figure 33 – The Spy technique in S-EM (LIU et al., 2002) applied to the purpose-oriented method.....	169
Figure 34 – Example of the <i>alignment of case CVs by ranking</i> .....	170
Figure 35 – The dataset of <i>positive</i> and <i>negative</i> instances to apply <i>feature weighting</i> algorithms.....	171
Figure 36 – Stage 3: Classifying candidate researchers as <i>fit</i> or <i>unfit</i> for the purpose .....	172
Figure 37 – The set of cases (C) and set of new cases (Q).....	173
Figure 38 – The neighbour ( $NN_k$ ) between a new case (Q) and cases (C) .....	174
Figure 39 – The two experiments to demonstrate the <i>purpose-oriented</i> <i>method</i> .....	177
Figure 40 – Accuracy by classifier and by job .....	181
Figure 41 – The set of example CVs and set of candidate CVs by regions of Brazil .....	185
Figure 42– The set of candidate CVs from the 2011 to 2015 .....	186
Figure 43– The data representation of a candidate researchers’ career trajectory .....	187
Figure 44 – The resulting datasets of Experiment II – Part I .....	189
Figure 45 – The process of initialization of the balanced set (M).....	190
Figure 46 – The alignment between the example CVs .....	191
Figure 47 – The purpose represented through weights .....	192
Figure 48 – The fit candidates distributed in the CNPq knowledge areas and the five regions of Brazil. ....	195
Figure 49 – Clusters of similarity between the members of Fiocruz MERG .....	196
Figure 50 – The rank of <i>fit</i> candidates to cluster-1, cluster-2 and cluster- 3.....	197
Figure 51 – The resulting datasets of Experiment II – Part II.....	201
Figure 52 – The resulting rank clusters for each year the target interval .....	202
Figure 53 – The evolution of similarity score of the 10 most similar fit candidates for the purpose of the assessment .....	206
Figure 54 – The distribution of fit candidates of Experiment II, Part-I and Part-II, by regions of Brazil.....	210
Figure 55 – The distribution of fit candidates of Experiment II, Part-I and Part-II, in the CNPq knowledge areas. ....	210

## LIST OF FRAMES

Frame 1 – Studies from PPGE GC/UFSC selected to this thesis .....	48
Frame 2 – Terms and concepts associated to research assessment .....	53
Frame 3 – Performance metrics applied to research assessment .....	59
Frame 4 – Critical aspects concerning to the use of research metrics ..	60
Frame 5 – Summary of the section 2.2 .....	66
Frame 6 – The early techniques to measure research collaboration .....	75
Frame 7 – Indicators and research collaboration methods based on the co-authorship principle .....	76
Frame 8 – Statistical properties of co-authorship networks based on Newman’s studies .....	77
Frame 9 – Summary of the section 2.3 .....	83
Frame 10 – Conceptual characteristics of cooperation, collaboration and coproduction .....	84
Frame 11 – Summary of the section 2.4 .....	92
Frame 12 – The use of CBR methodology on related studies and applications .....	105
Frame 13 – Common statistical measures used in machine learning ..	114
Frame 14 – Examples of normalization techniques .....	116
Frame 15 – The confusion matrix for binary classification accuracy .	117
Frame 16 – Measures to compute accuracy in binary classifications..	117
Frame 17 – Summary of the section 2.5 .....	119
Frame 18 – List of attributes characterizing research collaborators as researchers .....	138
Frame 19 – List of attributes characterizing research collaborators from the perspective of her or his institutions .....	139
Frame 20 – List of attributes characterizing research collaborators from the perspective of a researcher’s affiliations .....	140
Frame 21 – List of attributes characterizing research collaborators from the perspective of researchers’ accomplishments .....	141
Frame 22 – List of attributes characterizing research collaborators from the perspective of the collaborative process delivered to achieve accomplishments .....	143
Frame 23 – List of attributes that characterize research collaborators from the perspective of their career trajectories. ....	144
Frame 24 – List of attributes selected from the <i>Lattes database</i> for this thesis .....	149
Frame 25 – List of accomplishment type identified in the <i>Lattes database</i> for this thesis .....	150

Frame 26 – Final list of attributes derived from the accomplishment types, classified as solo or collaborative .....	151
Frame 27 – The evolution of a <i>fit candidate</i> toward the research group’s purpose .....	207
Frame 28 – The classification of the <i>fit candidate</i> 7550, as <i>fit</i> or <i>unfit</i> , in the last five years of career trajectory .....	208
Frame 29 – The evolution of an <i>unfit candidate</i> in an interval of career trajectory .....	208
Frame 30 – The classification of the <i>candidate</i> 7859, as <i>fit</i> or <i>unfit</i> , in the last five years of career trajectory .....	209
Frame 31 – The rank of 100 most similar <i>fit candidates</i> for the purpose of the Experiment II – Part II .....	243

## LIST OF TABLES

Table 1 – Most frequent authors included in this study. ....	132
Table 2 – Journals included in this study .....	133
Table 3 – Most frequent keywords in publications included in this study .....	134
Table 4 – The case base (CB) represented by a feature-value pairs data structure .....	156
Table 5 – The case base of CVs represented by a feature-value pairs data structure .....	157
Table 6 – The <i>positive set (P)</i> represented by a <i>feature-value pairs</i> data structure .....	167
Table 7 – The <i>CBR</i> representation for the <i>purpose-oriented classifier</i> .....	172
Table 8 – The rank of similar cases to each new case .....	174
Table 9 – Number of <i>fit and unfit</i> researchers in each dataset.....	179
Table 10 – Similarity score between candidate CVs to collaborative purposes.....	181
Table 11 – The set of example CVs and set of candidate CVs by CNPq knowledge areas .....	184
Table 12 – The set of example CVs and set of candidate CVs by CNPq knowledge sub-areas .....	185
Table 13 – The resulting classification of the Experiment II – Part I. ....	194
Table 14 – The statistical results of cluster-1, cluster-2 and cluster-3 ranks .....	197
Table 15 – The resulting rank of <i>fit candidates</i> to the purpose of the assessment, in which the experiment does not considers career trajectory. ....	198
Table 16 – The resulting classification of the Experiment II – Part II ....	202
Table 17 – The final career trajectory rank .....	203
Table 18 – The results of Experiment II.....	203
Table 19 – Distribution of <i>unfit candidates</i> from 2011-2015.....	204
Table 20 – New researchers in some period of the target interval .....	204
Table 21 – <i>Unfit candidates</i> in some period of the target interval .....	205
Table 22 – <i>Fit candidates</i> in some period of the target interval.....	205
Table 23 – The 10 most similar <i>fit candidates</i> for the purpose of the assessment .....	206



## LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AIF	Author Impact Factor
Altmetrics	Alternative metrics
A&HCI	Arts & Humanities Citation Index
CAPES	Coordination for the Improvement of Higher Education Personnel
CBR	Case-based reasoning
CC	Collaborative Coefficient
CFS	Correlation Based Feature Selection algorithm
CI	Collaborative Index
CMS	Compact Muon Solenoid
CNPq	National Council for Scientific and Technological Development
CPqAM	Aggeu Magalhães Research Center
CV	Curriculum vitae
CVs	Curricula vitae
CVs	Curriculums vitae
DBSCAN	Density-based Spatial Clustering of Application with Noise
DC	Degree of collaboration
DGP	Directory of Research Groups
DORA	San Francisco Declaration on Research Assessment
DTV	Definite Typical Value
EASST	European Association for the Study of Science and Technology
EGC	Graduate Program in Engineering and Knowledge Management
EM	Expectation Maximization algorithm
FSS	Fractional Scientific Strength
FP	Fractional Productivity
GS	Google Scholar
HGP	The Human Genome Project
IBL	Instance-Based Learning algorithms
ICCBR2016	The Twenty-Forth International Conference on Case-Based Reasoning
ID3	Iterative Dichotomiser algorithm
IF	Impact factor
IG	Information Gain algorithm
JIF	Journal Impact Factor

KBS	Knowledge-based system
KE	Knowledge Engineering
K-NN	K-Nearest Neighbors
KM	Knowledge Management
LHC	Large Hadron Collider
LIBER	Association of European Research Libraries
LOOCV	Leave-One-Out Cross-Validation
MERG	Microcephaly Epidemic Research Group
ML	Machine Learning
MCTI	Ministry of Science, Technology and Innovation
MCTIC	Brazilian Minister of Science, Technology, Innovation and Communication
MTD	Most Typical Deviation
MTV	Most Typical Value
NB	Naïve Bayes
NN	Nearest Neighbor
OECD	Organisation for Economic Co-operation and Development
P	Productivity
PhD	Doctor of Philosophy Degree
REF	UK Research Excellence Framework
R&D	Research and Development
SCI	Science Citation Index
SNA	Social network analysis
SSCI	Social Sciences Citation Index
S&T	Science and Technology
ST&I	Science, Technology and Innovation
STI	Science, Technology and Innovation
STI2014	2014 International Conference on Science and Technology Indicators
STI2016	2016 International Conference on Science and Technology Indicators
UFSC	Santa Catarina Federal University
UK	United Kingdom
USA	United States of America
WW II	Second War
WWW	World Wide Web
SCI	Science Citation Index
WoS	Web of Science

## SUMMARY

<b>1</b>	<b>INTRODUCTION</b> .....	<b>37</b>
1.1	CONTEXTUALIZATION .....	37
1.2	RESEARCH PROBLEM .....	38
1.3	RESEARCH QUESTION .....	39
1.4	HYPOTHESIS/ASSUMPTIONS .....	40
1.5	RESEARCH GOALS .....	40
<b>1.5.1</b>	<b>General Goal</b> .....	<b>40</b>
<b>1.5.2</b>	<b>Specific Goals</b> .....	<b>40</b>
1.6	JUSTIFICATION FOR CONDUCTING THIS STUDY .....	40
<b>1.6.1</b>	<b>Relevance</b> .....	<b>40</b>
<b>1.6.2</b>	<b>Originality</b> .....	<b>42</b>
<b>1.6.3</b>	<b>Contributions</b> .....	<b>42</b>
1.7	METHODOLOGY .....	43
1.8	DELIMITATION OF THIS STUDY .....	44
1.9	ADHERENCE TO PPGE/C/UFSC .....	47
1.10	LAYOUT .....	49
<b>2</b>	<b>LITERATURE REVIEW</b> .....	<b>51</b>
2.1	INTRODUCTION OF THE CHAPTER .....	51
2.2	RESEARCH ASSESSMENT .....	52
<b>2.2.1</b>	<b>Terms and concepts associated to research assessment</b> .....	<b>52</b>
<b>2.2.2</b>	<b>Historical context of research assessment</b> .....	<b>54</b>
<b>2.2.3</b>	<b>Current methods applied on research assessment</b> .....	<b>56</b>
<b>2.2.4</b>	<b>Performance metrics and indicators associated to research assessment methods</b> .....	<b>58</b>
<b>2.2.5</b>	<b>Critical aspects of the performance metrics on research assessment</b> .....	<b>60</b>
<b>2.2.6</b>	<b>Source of principles for assessing research quality</b> .....	<b>62</b>
<b>2.2.7</b>	<b>Future trends on research assessment.</b> .....	<b>64</b>
<b>2.2.8</b>	<b>Concluding remarks</b> .....	<b>66</b>

2.3	RESEARCH COLLABORATION .....	69
2.3.1	The historical context of research collaboration .....	69
2.3.2	The theoretical bases of research collaboration.....	71
2.3.3	Distinct meanings for the term “collaboration” .....	73
2.3.4	Current methods and performance metrics applied on research collaboration.....	75
2.3.5	Source of principles for assessing research quality on collaboration .....	79
2.3.6	Related interdisciplinary collaborative projects.....	79
2.3.7	Future trends on research collaboration.....	81
2.3.8	Concluding remarks.....	82
2.4	SOURCES OF KNOWLEDGE FOR RESEARCH ASSESSMENT .....	86
2.4.1	Scientists and Researchers.....	87
2.4.2	Research collaborators.....	87
2.4.3	Research decision makers.....	87
2.4.4	Careers trajectories.....	88
2.4.5	Curriculum vitae .....	88
2.4.6	Research publications .....	89
2.4.7	Bibliographical databases.....	89
2.4.8	CV databases .....	90
2.4.9	Concluding remarks.....	91
2.5	KNOWLEDGE ENGINEERING .....	93
2.5.1	Knowledge-based systems (KBSs).....	94
2.5.2	The CommonKADS methodology.....	96
2.5.3	The Case-based reasoning (CBR) methodology .....	99
2.5.4	Machine Learning methods.....	106
2.5.5	Concluding remarks.....	118
3	METHODOLOGY .....	121
3.1	INTRODUCTION OF THE CHAPTER.....	121

3.2	METHODOLOGICAL PROCEDURES .....	122
3.3	CONCLUDING REMARKS .....	126
<b>4</b>	<b>CONCEPTUAL DATA MODEL FOR RESEARCH COLLABORATORS.....</b>	<b>129</b>
4.1	INTRODUCTION OF THE CHAPTER .....	129
4.2	SYSTEMATIC LITERATURE REVIEW .....	129
<b>4.2.1</b>	<b>Stage I: Planning the review .....</b>	<b>130</b>
<b>4.2.2</b>	<b>Stage II: Conducting the review.....</b>	<b>130</b>
<b>4.2.3</b>	<b>Stage III: Reporting the review .....</b>	<b>131</b>
4.3	THE CONCEPTUAL DATA MODEL.....	134
<b>4.3.1</b>	<b>Scope and limitations of the conceptual data model.....</b>	<b>136</b>
<b>4.3.2</b>	<b>The attributes of research collaborators .....</b>	<b>137</b>
4.4	CONCLUDING REMARKS .....	145
<b>5</b>	<b>THE DATA .....</b>	<b>147</b>
5.1	INTRODUCTION OF THE CHAPTER .....	147
5.2	THE DATA SOURCE.....	148
5.3	THE ATTRIBUTE SELECTION PROCESS .....	148
5.4	THE DATA EXTRACTION PROCESS.....	155
5.5	THE DATA REPRESENTATION PROCESS .....	156
5.6	CONCLUDING REMARKS .....	157
<b>6</b>	<b>THE PURPOSE-ORIENTED METHOD .....</b>	<b>159</b>
6.1	INTRODUCTION OF THE CHAPTER .....	159
6.2	THE PURPOSE-ORIENTED CLASSIFIER .....	161
<b>6.2.1</b>	<b>Stage 1: Describing the problem .....</b>	<b>163</b>
<b>6.2.2</b>	<b>Stage 2: Learning weights to represent purpose.....</b>	<b>164</b>
<b>6.2.3</b>	<b>Stage 3: Classifying candidate researchers as <i>fit</i> or <i>unfit</i> for the purpose .....</b>	<b>171</b>
<b>6.2.4</b>	<b>Stage 4: Ranking candidate researchers.....</b>	<b>176</b>
6.3	CONCLUDING REMARKS .....	176
<b>7</b>	<b>USEFULNESS OF THE METHOD .....</b>	<b>177</b>
7.1	INTRODUCTION OF THE CHAPTER .....	177

7.2	EXPERIMENT I .....	178
7.2.1	Introduction .....	178
7.2.2	Data.....	178
7.2.3	Methodology .....	180
7.2.4	Purpose-oriented or purpose-independent assessment? .....	180
7.2.5	Concluding remarks.....	182
7.3	EXPERIMENT II.....	182
7.3.1	Introduction .....	182
7.3.2	Data.....	183
7.3.3	Methodology .....	187
7.3.4	Experiment II – Part I.....	188
7.3.5	Experiment II – Part II.....	199
7.3.6	Analysis .....	203
7.3.7	Concluding Remarks.....	210
8	CONCLUSIONS.....	213
8.1	RESEARCH QUESTIONS .....	213
8.2	CONTRIBUTIONS.....	217
8.3	FUTURE WORKS .....	219
	REFERENCES.....	221
	APPENDIX A.....	243

# 1 INTRODUCTION

## 1.1 CONTEXTUALIZATION

As stated by Collins, Morgan and Patrinos (2003), “*Good science can only happen with good scientists*”. Despite this evident requirement, determining researcher quality has been posed as a great challenge to decision makers (VAN NOORDEN et al., 2013). Hence, there is an increasing demand for efficient methods to assess researcher quality for purposes such as, recruitment, promotion, and grant awarding decisions (GARFIELD; MALIN, 1968; HAUSTEIN; LARIVIÈRE, 2015; LANE, 2010; LANE et al., 2015).

The challenge of assessing researcher quality was intensified in the 1960s, in the quest for governance instruments to supply and guide research policy decisions through more objective criteria (OKUBO, 1997). This was a period of great incentives to scientific and technological advancement, in which many funding agencies were created to support them (NARIN; HAMILTON, 1996; OKUBO, 1997). However, it was also a period in which decision makers realized that science needed much more funding to tackle the needs of humanity, and hence, more objective criteria were required to guide research policy decision making (NARIN; HAMILTON, 1996).

The Science Citation Index (SCI) was the first bibliometric indicator developed for measuring researcher quality (GARFIELD, 1964). The concept behind the SCI is that “*the number of papers a man publishes does indeed provide some measure of his or her activity*” (GARFIELD; MALIN, 1968). It is also the object of a vast literature, which includes research collaborator quality assessment. For instance, the most widely used metric to measure research collaboration is the *co-authorship of scientific papers* (BEAVER; ROSEN, 1978). This metric incorporates the concept of research cooperation between co-authors by crediting contributions to each co-author, so that the number of contributions credited will reflect their quality (VINKLER, 1993). Another metric used is the *co-authorship network* (NEWMAN, 2001, 2004), which is based on the *Social Network Analysis* (SNA).

Nowadays, objective metrics have been increasingly used to quantify scientific quality (VAN NOORDEN, 2010), however, despite the advantages of adopting objective criteria by decision makers, criticisms regarding misapplication of such metrics became a debate between researchers. For instance, as said by David and Frangopol (2015,

p.2256), “*indicators designed to evaluate journals are wrongly used to evaluate individuals and/or groups, or vice versa*”.

In response to this criticism, a group of researchers proposed a set of principles called the *Leiden Manifesto for research metrics* (HICKS et al., 2015). The Manifesto starts by recommending the use of objective metrics in support of qualitative judgement. It suggests aligning metrics to the mission and purposes of institutions, and emphasizes the need to acknowledge local instead of universal research. It states that measures should be transparent, in that, data and analyses can be verifiable. It stresses the importance of reviewing a researcher’s portfolio, and warns about the risk of unfair results when specifying information at distinct levels of abstraction. It also warns about researchers directing their work based on the metric and not based on broader impacts, and at the end, it emphasizes the importance of data quality and recency.

Bozeman, Fay and Slade (2013) reviews the literature on research collaboration, and suggests that more attention should be given to some research gaps, by proposing a research agenda for future studies. One of these research gaps is the lack of methods that go beyond the SCI to measure research collaboration. Considering that such methods are based on citation analysis, these authors recommend that “*collaboration research must find a better way to measure the impact to fundamental knowledge beyond citation rates*” (BOZEMAN; FAY; SLADE, 2013).

## 1.2 RESEARCH PROBLEM

This thesis investigates the problem of assessing researcher quality, in special the assessment of research collaborators in selection processes such as recruitment, promotion, and funding.

Considering the needs of governance instruments to supply and guide research policy decisions addressing researcher quality, the principles stated in the *Leiden Manifesto for research metrics* (HICKS et al., 2015), and the gap identified by Bozeman, Fay and Slade (2013), this thesis aims to answer the following research question “*How to assess researcher quality for collaborative purposes?*”.

Thus, this investigation requires initially to exploring four essential concepts: *Research assessment*, *research collaboration*, *research collaborators*, and *quality*.

Research assessment “includes the evaluation of research quality and measurements of research inputs, outputs and impacts, and embraces both qualitative and quantitative methodologies” (MOED, 2011)

Research collaboration, in a classical concept, is defined as “*the working together of researchers to achieve the common goal of producing new scientific knowledge*” (KATZ; MARTIN, 1997).

Research collaborators, as suggested by Katz and Martin (1997), are researchers who work together to advance scientific knowledge in research projects, scientific papers, or some other key step of scientific research.

*Quality is fitness for purpose* (JURAN; GODFREY, 1999), which is an definition that conceptualizes quality as dependent on perspectives, needs, and priorities of users, in this, it varies across user-groups.

Furthermore, Knowledge Engineering (KE) approaches are also investigated to propose a suitable solution to this research problem. *Knowledge Engineering* (KE) is a field of Artificial Intelligence (IA), which addresses the construction of knowledge-based systems (KBSs) based on knowledge modelling and knowledge representation (SCHREIBER et al., 2000). Among the approaches of the KE to be explored, *Case-Based Reasoning* (CBR) methodology has demonstrated significant results in both analytical or synthetical tasks associated to cognitive processes including analogical reasoning. Concerning the researchers’ data, *CV databases*, in particular the *Brazilian Lattes database* (lattes.cnpq.br), have motivated a series of studies on researchers, by providing high-quality data (LANE, 2010) and accurate information (PERLIN et al., 2017).

In the following sections, research questions, general and specific goals, justification, and delimitations for conducting this study will be presented in order to contextualize the whole study. In addition, the adherence of this thesis to the Graduate Program in Engineering and Knowledge Management at the Federal University of Santa Catarina (PPGEGC/UFSC) is evidenced. At the end, the structure of the document is outlined.

### 1.3 RESEARCH QUESTION

The issue that this thesis wants to address is in the understanding and answering to the following research question: “*How to assess researcher quality for collaborative purposes?*”. This question is supported by two sub-questions:

RQ1: How to conceptualize a data model to assess researcher quality with emphasis on research collaborators?

RQ 2: How to assess researcher quality?

## 1.4 HYPOTHESIS/ASSUMPTIONS

In order to answer the second sub-question, “RQ2: How to assess researcher quality?”, this study presents the hypothesis that “H2.1: A purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method”. In addition, the study is based on the assumption that “A2.1: Incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment”.

## 1.5 RESEARCH GOALS

### 1.5.1 General Goal

This work has as a general goal to propose a method to assess researcher quality for collaborative purposes.

### 1.5.2 Specific Goals

In order to achieve the general goal of this work, and based on the research questions, hypothesis and assumption previously stated, the following specific goals are proposed to guide the approach of this study.

1. Identify factors, concepts and elements suitable for research and collaboration assessment.
2. Identify methods and techniques of knowledge engineering, which can be used to implement approaches oriented to purpose for researcher quality assessment.
3. Develop a method capable to contribute with the researcher quality assessment, particularly on collaborative purposes.

## 1.6 JUSTIFICATION FOR CONDUCTING THIS STUDY

### 1.6.1 Relevance

The relevance of this study is backed by a literature review. I begin by pointing out the consensus that science has become deliberately more collaborative, as observed by Derek de Solla Price in his book “*Little Science, Big Science*” (PRICE, 1963), and fifty years later by Jonathan Adams, in his study “*Collaborations: the fourth age of research*” (ADAMS, 2013). Research collaboration has become crucial

for productivity in science, and its theoretical basis is related to the professionalization of science (BEAVER; ROSEN, 1978). For instance, the collaboration between researchers pushed them toward a global science, and hence, to the exponential growth in the number of scientific publications (HENNEMANN; RYBSKI; LIEFNER, 2012).

This global science, through the emergence of the Internet and the Web, supplied the cyberinfrastructure which made solving the grand challenges of science possible (OMENN, 2006), for example, the *Human Genome Project* (COLLINS; MORGAN; PATRINOS, 2003); the CMS experiment (CMS COLLABORATION, 2008); and the Brazilian Microcephaly Epidemic Research Group (MERC, 2016). These three examples are characterized as “*Big Science*” (PRICE, 1963), and they concerns to large collaborative projects that involve multiple investigators and conceptualizations of research problems from different institutions and cultures (COLLINS; MORGAN; PATRINOS, 2003; WELSH; JIROTKA; GAVAGHAN, 2006).

However, this global science demands instruments of research governance to regulate, coordinate, and monitor funded research activities, in order to conduct “*Big Science*” projects (CUMMINGS; KIESLER, 2011). According to Okubo (1997), the needs for such instruments became evident after the II World War, when a substantial increase in scientific production occurred, and the consequential increase in financial support demanded more efficient methods for research governance. Thus, the assessment of academic performance for selection, hiring, and funding decisions, has become a practice in funding agencies, universities, and public and private research institutes (HAUSTEIN; LARIVIÈRE, 2015; LANE, 2010).

The development of the Bibliometric field of study originated from this demand, with the goal of measuring science through citation analysis (GARFIELD, 1963). Over time, bibliographical databases and integrated information systems have been proposed, intensifying support on decisions about science (ABRAMO; D’ANGELO, 2011; HICKS et al., 2015).

Besides this current scenario of opportunities for investigation on more efficient methods to conduct global science, the OECD (2016) highlights the development of a new culture of collaborative awareness, which motivates research practices based on new paradigms such as “*open science*”, “*open access*”, “*open data*”, and “*open collaboration*”. Furthermore, this study promotes the idea that in Bibliometrics, research should be assessed on its own merits, as recommended by Hicks et al. (2015) in the Leiden Manifesto for research metrics.

In sum, this thesis proposes an approach to assess researcher quality for collaborative purposes, and its relevance is in its intrinsic adherence to the context described above. In the following subsections, I will justify the originality of the approach.

### 1.6.2 Originality

The originality of this thesis is specially in addressing research collaborators as units of analysis<sup>1</sup> in assessment processes, such as, selection, hiring, and funding decisions. This thesis focuses particularly on approaches that take into account the purpose of assessment to evaluate the quality of individual researchers in relation to their set of collaborative accomplishments.

Therefore, the proposed approach differs from co-authorship networks (NEWMAN, 2001, 2004) approaches, which are based on *Social network analysis* (SNA), in that, it considers the relationship patterns between the entire set of collaborators. In contrast to co-authorship networks, this proposed approach focuses on measuring the quality of individual collaborators.

This study is also original, in that, it is aligned to the concept that quality is *fitness for purpose* (JURAN; GODFREY, 1999). *Fitness for purpose* relies on the fact that quality depends on perspectives, needs and priorities of users. Furthermore, it is aligned to the principles of the Leiden Manifesto (HICKS et al., 2015), which recommends attention to purpose, context, and transparency, when creating research metrics. Thus, the approach characterizes the purpose of the assessment before effectively applying a method to calculate the researcher quality.

Finally, this study is unique in that it automatically assesses researcher quality, by analyzing the similarity between successful researchers and candidate researchers to a target selection process, through their curriculum vitae.

### 1.6.3 Contributions

This subsection will be focus particularly on four relevant contributions of this thesis for decisions in science and technology, in selection processes of recruitment, promotion, and funding.

---

<sup>1</sup> Unit of analysis: “A unit of analysis is the most basic element of a scientific research project. That is, it is the subject (the who or what) of study about which an analyst may generalize” (LEWIS-BECK; BRYMAN; LIAO, 2004).

The first contribution addresses the creation of a *purpose-oriented method* to assess researcher collaborators quality, based on their career trajectories. Thus, this thesis contributes providing to decision makers, a methodology or “*how to do*” to assess research collaborators based on their career trajectories, and the collaborative purpose of the assessment. Moreover, the *purpose-oriented method* assesses not only research collaborators, but researchers in general.

The second contribution concerns the efficiency of the *purpose-oriented method* to support qualitative expert assessments and guide research policy. For example, the proposed method allows to compare automatically large numbers of researchers’ curriculum vitae, analyzing each year of their career trajectories, in a target interval. In this analysis, similarity measures are applied, which contrasts the experience of successful researchers with candidate researchers.

The third contribution stresses the ability of the *purpose-oriented method* of representing the criteria of the assessment through the relative relevance of the attributes of the set of example CVs. This fact allows decision makers to enhance the comprehension about the collaborative purpose of the assessment. For example, analyzing the set of example CVs makes possible to understand different clusters of researchers, as well as, the role of individual researchers in collaborative purposes.

Such contributions are synthesizing in this sub-section. However, they will be presented in more details in Chapter 8 – Conclusions.

## 1.7 METHODOLOGY

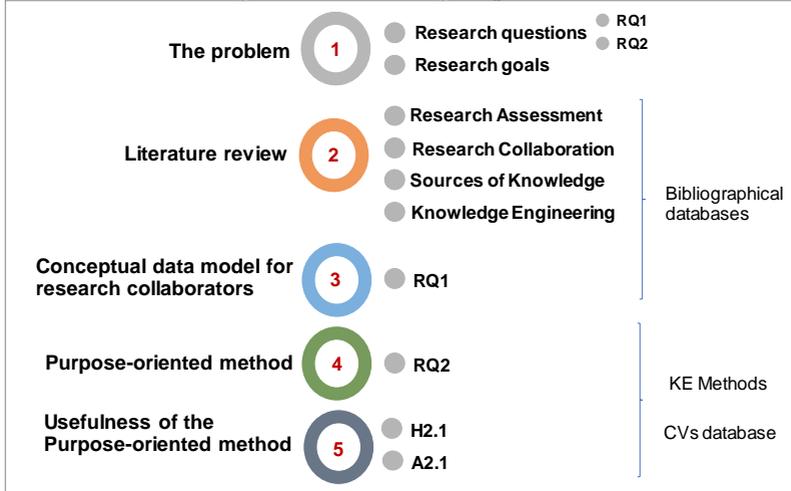
This thesis relies in a *quantitative world view*, in which, variables can be measured by instruments, and data can be analyzed using statistical procedures (CRESWELL, 2009). It is characterized as an *applied research*, which seeks for specific, and practical objectives, by applying the results in practical problems (OECD, 2015b). Furthermore, it focuses on a approaches based on interdisciplinary collaborations, which is characterized by the collaboration between researchers from different disciplines, with a common methodological approach and a shared problem (MOBJÖRK 2010).

The study is conducted in five steps, as illustrated in Figure 1. In step 1, the problem is defined, and the research questions are formulated;

In step 2, the literature is reviewed in order to scrutinize four constructs: Research assessment, research collaboration, sources of knowledge for research assessment, and Knowledge Engineering.

In Step 3 a systematic literature review, based on Tranfield et al (2003) is conducted to investigate the domain knowledge about researchers and research collaborators. Este resulting domain knowledge is represented by the conceptual data model for research collaborators.

Figure 1 – Methodological procedure



Source: The author, 2017.

In step 4, Knowledge Engineering methodologies are used to design the proposed purpose-oriented method. This step includes the specification of the data used in this study. At last, in the step 5, the usefulness of the proposed method is demonstrated through two experiments. Details of this methodological procedures will be described in Chapter 3.

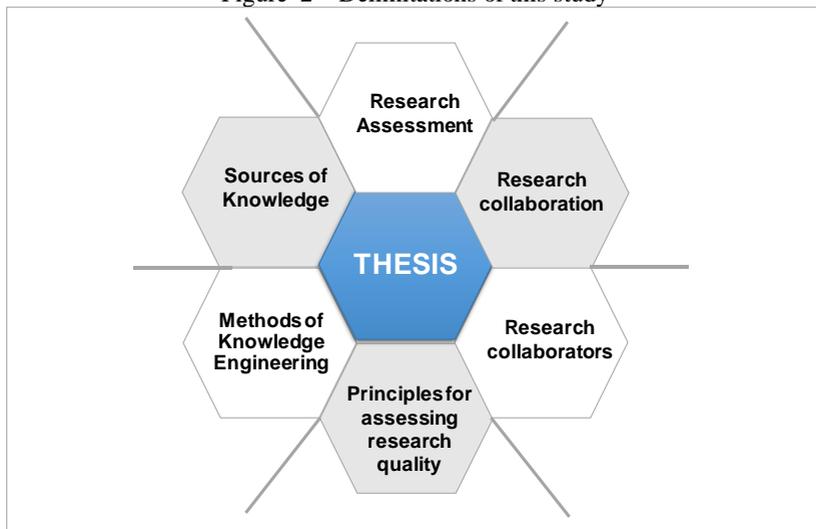
## 1.8 DELIMITATION OF THIS STUDY

This section presents the scope of this thesis, addressing its limits regarding (i) research assessment, (ii) research collaboration, (iii) research collaborators, (iv) principles for assessing research quality, (v) methods of knowledge engineering, and (vi) sources of knowledge, as illustrated in Figure 2.

*Research assessment* is an instrument of governance used by decision makers to foment excellence on conducting scientific research. It measures scientific inputs, outputs and impacts, to guide future research

decisions (LANE et al., 2015; MOED, 2011; PATRICK; STANLEY, 1996; VASILEIADOU, 2011). This thesis emphasizes research assessment, particularly focusing on methods, metrics, and models of measurement.

Figure 2 – Delimitations of this study



Source: The author, 2017.

*Research collaboration* is defined as “*the working together of researchers to achieve the common goal of producing new scientific knowledge*” (KATZ; MARTIN, 1997, p. 7). This thesis relies in this concept, and addresses the problem of assessing the quality of researchers who work in scientific collaboration. The focus on research collaboration was previously justified in the Section 1.4, which emphasized that research collaboration is one of the key elements in the professionalization of science and has become crucial for its growth. As limitations of the scope, this thesis does not address aspects of coproduction and transdisciplinary collaboration, such as the collaboration of researchers in projects outside academia. Furthermore, this thesis does not investigate research evaluation processes as they are executed in universities, funding agencies, and public and private institutions.

Research collaborators as suggested by Katz and Martin (1997), are researchers who work together to advance scientific knowledge in research projects, scientific papers, or some other key step of scientific

research. It is important to point out that this thesis is not limited to studies on research collaborators. It investigates researchers in general, however it focuses particularly on the research collaborator in the assessment processes, such as, selection, hiring, and funding decisions. In this sense, this thesis differs from studies on *co-authorship networks* (NEWMAN, 2001, 2004), which identify relationship patterns between individuals, and emphasizes the entire set of collaborators. Furthermore, this thesis does not address the investigation of subjective aspects, for instance, those studied in cognitive science and psychology; the social context of research collaborators, as well as, their tacit knowledge.

As aforementioned, for assessing research quality this thesis considers that quality is fitness for purpose (JURAN; GODFREY, 1999), as well as, it take into account the 10 principles of the Leiden Manifesto for research metrics (HICKS et al., 2015). The Leiden Manifesto has motivated a series of studies since it was presented in the STI2014 conference. It gathers, in only 10 principles, many of the ideas proposed by the other studies, such as, The Metric Tide (WILSDON et al., 2015) and DORA (AMERICAN SOCIETY FOR CELL BIOLOGY, 2015). In general, researchers have agreed to the relevance of the Manifesto and recommend its application on Altmetrics (BORNMANN; HAUNSCHILD, 2016), its adoption in libraries (COOMBS; PETERS, 2017), and also as a trend in research metrics (OECD, 2016).

This thesis applies methods of knowledge engineering to design the *purpose-oriented method*, which proposes assessing researcher quality taking into account their scholarly experiences along their career trajectories. To this end, *Case-based reasoning* (CBR) is adopted as a methodology to implement the propose method (e.g., RICHTER; WEBER, 2013). CBR was chosen because it is associated to cognitive processes such as analogical reasoning (MÁNTARAS et al., 2005), in that, it retrieved from memory past experiences that can be appropriately transformed, applied to the new situation, and stored for future use (CARBONELL; MICHALSKI; MITCHELL, 1983).

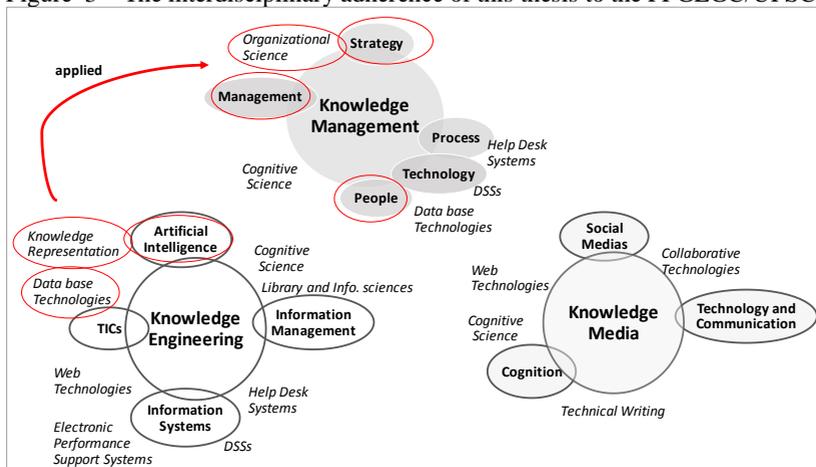
At last, this thesis adopts the curriculum vitae (CV) as source of knowledge. In particular, the *Brazilian Lattes database* (lattes.cnpq.br) was chosen as data source, due to its wealth of information, and openness to all S&T Brazilian institutions, allowing me to demonstrate the proposed method.

## 1.9 ADHERENCE TO PPGEGC/UFSC

The Graduate Program in Engineering and Knowledge Management of Federal University of Santa Catarina (PPGEGC/UFSC) has its institutional identity established on three interdisciplinarity subfields: The first is *Knowledge Engineering* (KE), a discipline originated from artificial intelligence (AI) that deals with modeling and representation of knowledge. The second is *Knowledge Management* (KM), which is based on disciplines such as, information management, strategy, and competitive intelligence, among others; KM provides methodologies to measure intangible assets of organizations. The third is *Knowledge Medias* that has the goal to disseminate knowledge in the organizations, and includes disciplines such as philosophy of science, epistemology of journalism and technological media, (PACHECO, 2010).

This thesis concerns Knowledge Engineering applied to organizations, with particular attention to research collaboration in institutions of science and technology. The adherence of the thesis to PPGEGC/UFSC is illustrated through Figure 3, which shows the three interdisciplinarity areas and respective subfields. In order to evidence the relation between the thesis and the PPGEGC/UFSC, the subfields addressed in the course of this study are highlighted.

Figure 3 – The interdisciplinary adherence of this thesis to the PPGEGC/UFSC



Source: Adapted of Pacheco et al. (2013)

Figure 3 emphasizes that this thesis proposes applying KE (i.e., disciplines of artificial intelligence, knowledge representation and database technologies) to solve problems of KM (i.e., problems related to strategy in research organizations).

Given this interdisciplinary adherence, I searched the PPGEGC/UFSC database (<http://btd.egc.ufsc.br/>) for studies closest to my work, on the topics of collaboration, research assessment and KE methods. After reading titles and abstracts of 190 theses, 11 of them were selected from the three areas of PPGEGC/UFSC, which are listed in Frame 1.

Frame 1 – Studies from PPGEGC/UFSC selected to this thesis

Source	Title	PPGEGC Area
(MARQUES, 2016)	<i>Reforming technology company incentive programs for achieving knowledge-based economic development: A Brazil-Australia comparative study</i>	Knowledge Management
(TAXWEILER, 2016)	<i>Um Modelo Para a Extração de Perfil de Especialista Aplicado às Ferramentas de Expertise Location e Apoio à Gestão do Conhecimento.</i>	Knowledge Engineering
(MANHÃES, 2015)	<i>Innovativeness and prejudice: designing a landscape of diversity for knowledge creation.</i>	Knowledge Management
(CECI, 2015)	<i>“Um modelo baseado em casos e ontologia para apoio à tarefa intensiva em conhecimento de classificação com foco na análise de sentimentos”</i>	Knowledge Engineering
(BORDIN, 2015)	<i>” Framework baseado em conhecimento para análise de rede de colaboração científica”</i>	Knowledge Engineering
(BRAGLIA, 2014)	<i>“Um Modelo Baseado em Ontologia e Extração de Informação como Suporte ao Processo de Design Instrucional na Geração de Mídias do Conhecimento”</i>	Knowledge Medias
(SALM JUNIOR, 2012)	<i>” Padrão de projeto de ontologias para inclusão de referências do novo serviço público em plataformas de governo aberto”</i>	Knowledge Engineering

(cont.)

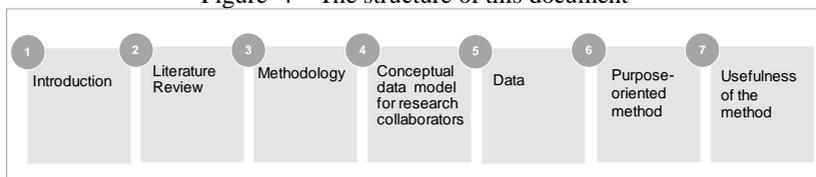
Source	Title	PPGEGC Area
(SARTORI, 2011)	" Governança em agentes de fomento dos sistemas regionais de CT&I"	Knowledge Engineering
(BOVO, 2011)	"Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais"	Knowledge Engineering
(BALANCIERI, 2010)	" Um método baseado em ontologias para explicitação de conhecimento derivado da análise de redes sociais de um domínio de aplicação"	Knowledge Engineering
(RIBEIRO-JÚNIOR, 2010)	" Modelo de sistema baseado em conhecimento para apoiar processos de tomada de decisão em ciência e tecnologia"	Knowledge Engineering
(RAUTENBERG, 2009)	"Modelo de conhecimento para mapeamento de Instrumentos da gestão do conhecimento e de agentes computacionais da engenharia do conhecimento baseado em ontologias"	Knowledge Engineering
(SOUZA, 2009)	"Gestão das Universidades Federais brasileiras: uma abordagem fundamentada na Gestão do Conhecimento"	Knowledge Management

Source: The author, 2017.

## 1.10 LAYOUT

This document is structured in seven chapters, as presented in Figure 4.

Figure 4 – The structure of this document



Source: The author, 2017.

Chapter 1 is this introduction, in which the research problem to be solved is stated. The chapter includes the research questions, goals,

justification, scope and limitations, adherence to PPGE/C/UFSC, and this layout, which outlines the structure of this document.

Chapter 2 reviews the literature providing the state of the art. It supports the whole study, and it is organized through four key constructs: Research assessment, Research collaboration, Sources of knowledge, and Knowledge Engineering.

Chapter 3 outlines the methodological procedures used to develop this study.

Chapter 4 investigates the domain knowledge on research collaborators, and presents it in a conceptual data model used to assess researcher quality. The relevance of this chapter is in gathering from the literature a set of metadata on researchers, and organize them in a data model that will be used as input to the proposed method.

Chapter 5 describes the data used in this thesis, presenting the data source, the process of attribute selection, the data extraction, and how these data will be represented in the proposed method.

Chapter 6 introduces the purpose-oriented method to assess researcher quality for collaborative purposes. The proposed method consists of four steps: Step 1 describes the problem; Step 2 learns weights to represent the purpose of the assessment; Step 3 classifies candidate researchers as fit or unfit for the purpose of the assessment; and Step 4 ranks the candidate researchers classified as fit for the purpose of the assessment.

Chapter 7 demonstrates the usefulness of the proposed method through application scenarios concerning assessment processes in science and technology (S&T).

## 2 LITERATURE REVIEW

### 2.1 INTRODUCTION OF THE CHAPTER

Chapter 2 reviews the literature on the four main constructs of this thesis, *research assessment*, *research collaboration*, *sources of knowledge for research assessment*, and *knowledge engineering*, as shown in Figure 5.

The chapter starts by reviewing the state of the art on research assessment focusing on its main elements (i.e., concepts, methods, metrics, critical aspects, recommendations, and trends).

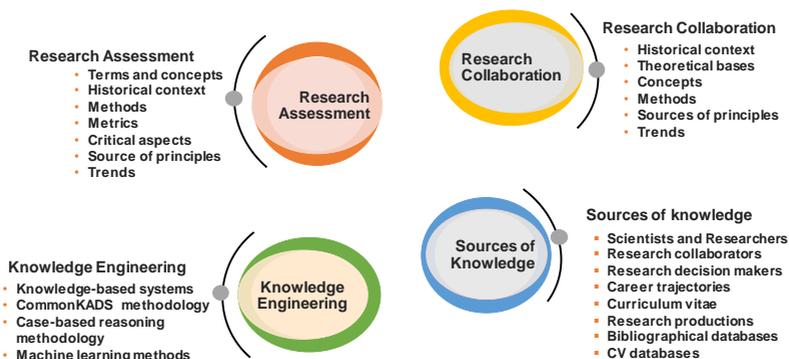
The second section explores research collaboration emphasizing its roots, theoretical bases, epistemological concepts, methods, metrics, and trends.

The third section, sources of knowledge for research assessment, considers for this thesis, the following sources of knowledge: researchers, research collaborators, career trajectories, curriculum vitae, research productions, bibliographical databases, and CV databases.

The fourth section, knowledge engineering, examines the concept of knowledge-based systems (KBSs), as well as, methodologies and methods based on artificial intelligence that are used to construct KBSs.

At the end of each section, a concluding remarks section highlights their main points and contributions.

Figure 5 – The four literature review constructs for this thesis



Source: The author, 2017.

## 2.2 RESEARCH ASSESSMENT

Over time, scientists have faced “*grand challenges*” in their search to understand and explain how the universe operates (OMENN, 2006). Such challenges are particularly visible through the technological achievements of modern society after the mid-20th century, which have contributed to the expansion of scientific inquiry and hence changed the way that scientists work (BREMBS; BUTTON; MUNAFÒ, 2013; CUMMINGS; KIESLER, 2014; WALTON; ZHANG, 2013).

As said by Suresh (2012, p.337), “*global challenges need global solutions*”. This statement emphasizes scientists’ demand for new methods and practices of research, especially those able to help them to accelerate their discoveries and connections (COLLINS, 2010). In the same way, policy makers need instruments of governance to support these great challenges in conducting scientific research, assessing their impacts, and guiding future funding decisions (LANE, 2010; LANE et al., 2015; LARGENT; LANE, 2012; STEEL et al., 2004; VASILEIADOU, 2011). For example, Lane (2010, p.488) emphasizes the need “*to make a science metric more scientific*”, and Moed (2011) highlights the role of research assessment for decision makers, in the governance of scientific-scholarly research.

This literature review explores studies particularly on research assessment, which involves the fields of history of science, research policy, research quality, and governance of science. The investigation of these fields is essential because research assessment includes, according to Moed (2011), “*the evaluation of research quality and measurements of research inputs, outputs and impacts, and embraces both qualitative and quantitative methodologies*”. As results, the next sub-sections present elements associated to the term “*research assessment*”. They briefly describe concepts, the historical context, current methods and performance metrics, as well as, criticisms and principles for assessing research quality. In addition to this context, future trends for research assessment on the perspective of OECD (2016) end this review.

### 2.2.1 Terms and concepts associated to research assessment

In this section, we introduce the most basic concepts to the study of research assessment, having as initial point, the definition of Moed (2011), as mentioned above.

Frame 2 presents the root of terms and concepts associated to “*research assessment*”, identified on dictionaries.

Frame 2 – Terms and concepts associated to research assessment

<b>Term</b>	<b>Concept</b>	<b>Source</b>
Science	<i>"What is known, knowledge (of something) acquired by study"; "experiential knowledge"; "collective human knowledge, especially that gained by systematic observation, experiment, and reasoning".</i>	Online Etymology Dictionary
	<i>"knowledge about or study of the natural world based on facts learned through experiments and observation"</i>	Merriam Webster Dictionary
Scientist	<i>"1834, a hybrid coined from Latin scientia (see science) by William Whewell (1794-1866), by analogy with artist"</i>	Online Etymology Dictionary
Research	<i>"Scientific inquiry"</i>	Online Etymology Dictionary
	<i>"careful study that is done to find and report new knowledge about something"; "careful or diligent search"</i>	Merriam Webster Dictionary
Quality	<i>"degree of goodness"; "Sense of "be fit for a job"</i>	Online Etymology Dictionary
	<i>"how good or bad something is"</i>	Merriam Webster Dictionary
Assess / Assessment	<i>"To estimate the value of property for the purpose of taxing it"; "to judge the value of a person, idea, etc."</i>	Online Etymology Dictionary
	<i>"to determine the importance, size, or value of"</i>	Merriam Webster Dictionary
Appraisal	<i>"the act of examining someone or something in order to judge their qualities, success, or needs"</i>	Cambridge Dictionary
Metric	<i>"pertaining to the system of measures based on the meter"</i>	Online Etymology Dictionary
Measure	<i>Action in "that to which something is compared to determine its quantity"</i>	Online Etymology Dictionary

(cont.)

Term	Concept	Source
Evaluation	<i>"action of appraising or valuing, "to find the value of";</i>	Online Etymology Dictionary
	<i>to judge the value or condition of (someone or something) in a careful and thoughtful way</i>	Merriam Webster Dictionary
Estimation	<i>"process of forming an approximate notion"</i>	Online Etymology Dictionary
Context	<i>The interrelated conditions in which something exists or occurs</i>	Merriam Webster Dictionary
	<i>The situation within which something exists or happens, and that can help explain it</i>	Cambridge Dictionary
Purpose	<i>"intention, aim, goal, propose"</i>	Online Etymology Dictionary

Source: Adapted of the Online Etymology Dictionary<sup>2</sup>, the Merriam Webster Dictionary<sup>3</sup>, and the Cambridge Dictionary<sup>4</sup>.

## 2.2.2 Historical context of research assessment

The origin of research assessment is in the quest for better practices of research governance to achieve excellence, transparency and accountability on conducting scientific research (MOED, 2011; PATRICK; STANLEY, 1996; VASILEIADOU, 2011).

The first initiatives towards the professionalization of science dates back to the scientific revolution, between the 16<sup>th</sup> and 19<sup>th</sup> centuries, in which the bases of modern science were established through several events that have had strong influence over the practices of research assessment (BEAVER; ROSEN, 1978; STEEL; LACH; WARNER, 2009; WRAY, 2009). The scientific method and its quantitative aspects are significant contributions of the *Positivist movement* (STEEL; LACH; WARNER, 2009). It was also in those years that the term “*science*” started to be used (WRAY, 2009), and professional organizations were

<sup>2</sup> Online Etymology Dictionary: <http://www.etymonline.com>

<sup>3</sup> Merriam Webster Dictionary: <http://www.merriam-webster.com>

<sup>4</sup> Cambridge Dictionary: <http://dictionary.cambridge.org>

created to provide support to scientific research, such as, the Royal Society (1660), and the National Academy of Sciences in the United States (1863). Furthermore, it was in the British Association for the Advancement of Science that the term “*scientist*” was said for the first time, by William Whewell, in 1833 (SNYDER, 2011).

However, it was after the II World War, in the “*era of quantitative, computer-tabulated science metrics*” (VAN NOORDEN, 2010, p. 864) that the research assessment began effectively possible due to factors such as, Bibliometrics studies, the creation of the Science Citation Index (SCI) (GARFIELD, 1964), and posterior availability of online bibliometric databases (HAUSTEIN; LARIVIÈRE, 2015; VAN NOORDEN, 2010). This was a period of great incentives to science growth (e.g. the creation of funding agencies and the contribution of science to technological advancement) (NARIN; HAMILTON, 1996; OKUBO, 1997). In the other hand the funding to the scientific community was being limited (NARIN; HAMILTON, 1996). Thus, the scenario was of a substantial increase in scientific production, and consequently the financial support to it demanded more efficient methods than those based only on subjective judgement of researchers and their productions (BECK, 1978).

In 1970s and 1980s practices of governance started to be adopted by countries that focused on improving education and infrastructure (FRIEDMAN, 2005). Research governance is a decision-making processes related to conducting scientific research (VASILEIADOU, 2011), and in the same line, a research governance system can be understood as a set of mechanisms to regulate, coordinate, and monitor publicly funded research activities (WOELERT; MILLAR, 2013). In regards to scientific research, it became clear that science and technology were expensive and the environmental and social problems would take time to be solved (OKUBO, 1997). Thus, in this period the scientific research became more professional, for instance, research institutions created more efficient evaluation systems based on knowledge (OKUBO, 1997).

In 1990s, with the emergence of the Internet and the World Wide Web (WWW), “*The new age of connectivity*” (FRIEDMAN, 2005, p. 60) began. As a consequence of the technological advancement, the globalization of scientific work was intensified by the increasing interaction between the scientists, resulting in the growth of scientific collaboration (BROWN, 2000). Besides this, *e-science* and *cyberinfrastructure* made the work of scientists more productive (ATKINS et al., 2003; HEY; HEY, 2006).

After the year 2000, sophisticated technologies allowed the growth of electronic publication, and a wide access to integrated information systems, which provided decision makers a set of tools to manipulate indicators in bibliometric databases, and improve quality to research assessment processes, such recruitment, promotion, and grant awarding decisions (MOED, 2011; VAN NOORDEN, 2010).

In the last decade, decisions about science have been increasingly based on Bibliometrics (ABRAMO; D'ANGELO, 2011; HICKS et al., 2015). For instance, funding agencies are using metrics to allocate funds according to the past performance of researchers and institutions (LANE, 2010; MOED, 2011). Similarly, countries have used the number of papers published as an indicator of scientific performance to demonstrate their scientific capability (ALLIK, 2013). Even though there are advantages to research metrics, many criticisms have been made concerning their abusive use by decision makers (HAUSTEIN; LARIVIÈRE, 2015; HICKS et al., 2015; LANE, 2010; VAN NOORDEN, 2010).

### **2.2.3 Current methods applied on research assessment**

Peer-review is the most common and the oldest method to measure the researchers' quality and the results of their work (HAUSTEIN; LARIVIÈRE, 2015). The peer-review process has its origin in 1750s, on first journal published by the Royal Society, which a select group of members recommended the manuscripts received to the editor for publishing (SPIER, 2002). The approach involves asking experts, to express their professional opinions about the significance of the work, the validity of the methodology, the analysis and conclusions of the work, and the clarity and simplicity of its presentation (LAWANI; ROAD, 1986).

Bibliometrics came from the demand for more efficient methods to measure qualitative results of scientific research (BEAVER, 2012; LANE, 2013; OKUBO, 1997). It concerns "*The application of mathematics and statistical methods to books and other media of communication*" (PRITCHARD, 1969, p. 349), and it was firstly known as "statistical bibliography" (Hulme, 1923). The Bibliometric approach is based on "the notion that the essence of scientific research is the production of knowledge and that scientific literature is the constituent manifestation of that knowledge" (OKUBO, 1997).

Narin et al. (1994) describe three basic principles addressed by Bibliometrics. The first principle is associated to the activity measurement of simple counts of scientific production, such as articles

and patents. The second is related to the impact measurement, which counts the number of times that scientific productions (e.g., articles and patents) are cited. The third concerns to the linkage measurement, or the number of citations linking scientific productions (e.g., article to article, patent to patent, and article to patent).

Despite the studies in Bibliometrics not being new, until the mid-20<sup>th</sup> century their analysis were limited and difficult to compute (HICKS; MELKERS, 2012; OKUBO, 1997). One reason for their inefficiency was the still incipient computational systems, and the other was the absence of an adequate metric to measure science (SMITH, 2012). Then, as a result of the studies of the linguist Eugene Garfield, the Science Citation Index (SCI) was created (GARFIELD, 1964), and Bibliometrics became a practical tool for the evaluation of scientific production oriented towards science policy (HAUSTEIN; LARIVIÈRE, 2015; HICKS; MELKERS, 2012; OKUBO, 1997).

Simultaneously to bibliometric studies, Russian researchers also investigated quantitative methods for science, which they called “*scientometrics*” (GALYAVIEVA, 2013; MINGERS; LEYDESDORFF, 2015). Nalimov (1966) was who initially proposed the term “*scientometrics*”. In 1978, the international journal *Scientometrics* was launched with the goal of publishing papers concerning quantitative aspects of the “*science of science and science policy*” (BECK, 1978, p. 3). Thus, the journal *Scientometrics* became a representative discussion forum of contemporary bibliometric and scientometrics fields of studies (OKUBO, 1997), whose difference is in that scientometrics studies concern to the development of citation analysis (MINGERS; LEYDESDORFF, 2015).

Informetrics is another field associated to the studies on mathematical and statistical methods applied to the scientific communication process. It was first defined by Nacke (1979) as “*the study of the application of mathematical methods to the objects of information science to describe and analyze their properties, establish laws, and perform decision making*”. Informetrics makes the convergence between the areas of library science, sociology of science, history of science, science policy and information retrieval (GALYAVIEVA, 2013).

Altmetrics, or alternative metrics, is the most recent field concerning methods based on the scientific communication process, which was named as *Scientometrics 2.0* by *Priem (2010)*. Altmetrics gathers a set of scientometrics web-based metrics associated to environments such as social media, online reference managers, collaborative encyclopedias, blogs, scholarly social networks, and

conference organization sites (PRIEM; GROTH; TARABORELLI, 2012). These alternative metrics go beyond traditional citation analysis, and proposes measuring the scholarly impact using data from downloads, link indexes, and scholastic bookmarking (PRIEM, 2010; PRIEM et al., 2010)

#### **2.2.4 Performance metrics and indicators associated to research assessment methods**

The *Science Citation Index* (SCI) was the first bibliometric indicator developed for measuring scientific productivity (GARFIELD, 1964). It was initially created by the linguist Eugene Garfield in 1955 for purposes of scientific literature retrieval (GARFIELD, 1964), for instance, “*Given a specific published paper, one can find all subsequent papers that cite it by simply knowing the first author and year of the paper in question*” (GARFIELD; MALIN, 1968, p. 5). However, the SCI was officially launched 1964, as an instrument for analysis of research activity measurement, which is an author-based indicator composed of two parts (i.e., the author index and the citation index), to achieve two purposes: to identify the scientist’s publications and their respective citations (GARFIELD; MALIN, 1968).

The core concept of SCI is that the number of citations received by a paper reflects its influence, and consequently legitimizes the scientist’s authority and prestige (BEAVER, 2012; BORNMANN; HAUNSCHILD, 2015; HAUSTEIN; LARIVIÈRE, 2015). Furthermore, for Abramo and D’Angelo (2016, p. 680), “*publications represent scientific advances, and citations the related value*”.

The SCI was the root of a profusion of metrics applied to research assessment (VAN NOORDEN, 2010). For instance, the Social Sciences Citation Index (SSCI), in 1973, the Arts & Humanities Citation Index (A&HCI), since 1978, and the most widely used, the “Journal Impact Factor (JIF)” (GARFIELD, 2007; MINGERS; LEYDESDORFF, 2015; VAN NOORDEN, 2010). Moreover, the SCI favored the emergence of new generations of studies, authors and bibliometric indexes as listed in Frame 3.

Frame 3 – Performance metrics applied to research assessment

<b>Metric</b>	<b>Description</b>
Paper count	Paper count is the most basic bibliometric measure, and just count the number of papers of a researcher or institution (THOMSON-REUTERS, 2008).
Citation count	Citation count is the number of times a researcher or research paper is cited by others in some time period (THOMSON-REUTERS, 2008; VAN NOORDEN, 2010).
Impact Factor (IF)	IF is related to both journal and author impact, and describes the average number of citations per published paper in some time period (GARFIELD, 2007).
Journal Impact Factor (JIF)	JIF is based on the number of “ <i>cites</i> ” in the current year to any items published in the journal during the previous two years, divided by the number of articles published during the same two years (GARFIELD, 2007). It represents the frequency with which an average article in a journal is cited (VAN NOORDEN, 2010).
Author Impact Factor (AIF)	AIF was created by Pan and Fortunato (2014) as an extension of the IF to authors, which considers the papers of an author instead of the papers published in a journal.
Field baselines	Field baselines is computed by the average citations per paper, for papers in a field (THOMSON-REUTERS, 2008).
H-index	Hirsch (2005) proposed a single number (h) to characterize the scientific output of a researcher, combining the productivity (number of papers) and impact (number of citations) of a scientist.
Eigenfactor	Bergstrom (2007) created the Eigenfactor, which is based on Google PageRank (PAGE et al., 1998), however it takes citations in the academic literature excluding self-citations to identify the most influential journals.
Web-based metrics	The emergence of metrics based on Web 1.0 and Web 2.0 represent the new view of impact of scholarship, which measure scientific activities through Web pages, download, bookmarks, tweeter, or blogs (PRIEM, 2010; ROEMER; BORCHARDT, 2012). For example, the number of times a research paper is accessed or downloaded online (VAN NOORDEN, 2010)

Source: The author, 2017.

## 2.2.5 Critical aspects of the performance metrics on research assessment

Despite the vast literature and potential use of performance metrics based on Bibliometrics, there are problematic issues in respect to the misapplication use of indicators in the evaluation process. To evidence this debate, I categorized 12 criticisms found in nine publications, present in this literature review, in seven critical aspects: the general use of research metrics, purpose, context, transparency and flexibility, data quality, time in metrics, and normalization. Frame 4 lists these critical aspects.

Frame 4 – Critical aspects concerning to the use of research metrics

Aspect	Criticism	Source
General use of research metrics	<i>“Scientometric data have sometimes been used in inappropriate ways ... For example, indicators designed to evaluate journals are wrongly used to evaluate individuals and/or groups, or vice versa”.</i>	(DAVID; FRANGOPOL, 2015, p. 2256)
General use of research metrics	<i>“Part of the debate about the impact factor is not so much about the indicator itself but more about the way in which the indicator is used for research assessment purposes”.</i>	(WALTMAN, 2016, p. 381).
General use of research metrics	<i>“Scientific performance indicators are proliferating — leading researchers to ask afresh what they are measuring and why”</i>	(VAN NOORDEN, 2010, p. 864)
General use of research metrics	<i>“The measurement of scientific activity should be more scientific” and “It should focus on scientists, networks of scientists (the doers of science), and their subsequent activities not just the documents created by the scientists”.</i>	(LANE, 2013, p. 13)
Purpose	<i>“...what are the objectives that we have understood as our reference? The short and long-term objectives of individual researchers or research institutions can be different, and more than one; they can vary over time, and within and between countries, and have different and varying importance”.</i>	(ABRAMO; D’ANGELO, 2016, p. 680)

(cont.)

<b>Aspect</b>	<b>Criticism</b>	<b>Source</b>
Context	<i>“Scientometricians can no longer merely be data providers or indicator builders. They need to be able to put the data in the right context”.</i>	STI 2014 Leiden – Preface
Context	<i>“Different types of indicators might be needed in different contexts”.</i>	STI2016, 3th Plenary Session
Context	<i>“How can we, for instance, use indicators to capture the performance of an organization against its research mission when these are peculiar to a local context?”</i>	STI2016, 5th Plenary Session
Transparence and flexibility	<i>“it should be feasible to create more reliable, more transparent and more flexible metrics of scientific performance”.</i>	(LANE, 2010, p. 489)
Data quality	<i>“Another aspect that is essential in bibliometric studies is the quality of data. This involves the selection of a suitable database and cleaning of bibliographic metadata”</i>	(HAUSTEIN; LARIVIÈRE, 2015, p. 5)
Time in metrics	<i>“The integration of time in some metrics, and its absence from others, can also have a huge impact on indicators and rankings. The difficulty in trying to incorporate time into a metric is that we need to understand its role within the information system”.</i>	(STUART, 2015, p. 849)
Normalization	<i>“Comparing the publication output and citation impact of authors, institutions, journals and countries without an accurate normalization is thus like comparing apples with oranges”.</i>	(HAUSTEIN; LARIVIÈRE, 2015, p. 6)
Normalization	<i>“Although normalized indicators are the best way to compare citation impact of different entities in a fair way, the complex structures of scholarly communication are difficult to capture in one indicator of citation impact”.</i>	(HAUSTEIN; LARIVIÈRE, 2015, p. 7)

Source: The author, 2016.

## 2.2.6 Source of principles for assessing research quality

Considering the critical aspects listed in the Frame 4 , I searched in the literature sources of principles to the appropriated use of performance metrics, and found four studies on this matter, which are: The Handbook on Constructing Composite Indicators (OECD, 2008), The San Francisco Declaration on Research Assessment (AMERICAN SOCIETY FOR CELL BIOLOGY, 2015), The Metric Tide (WILSDON et al., 2015), and The Leiden Manifesto for Research Metrics (HICKS et al., 2015).

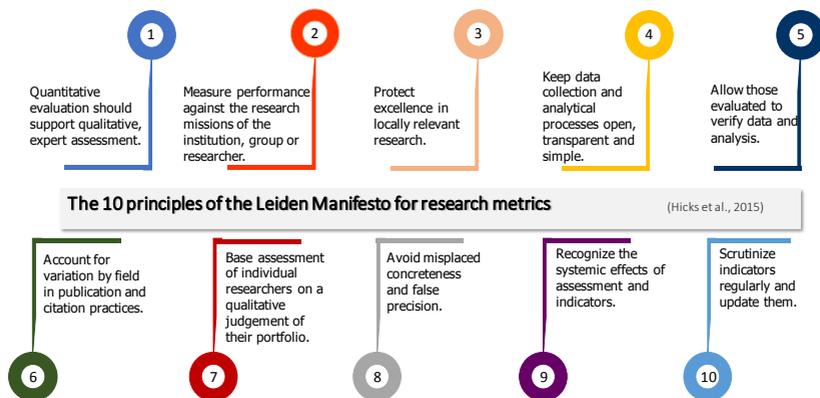
The Handbook on Constructing Composite Indicators, from the Organization for Economic Co-operation and Development (OECD), brings as the main recommendation, the inclusion of purpose in designing metrics, because they rely on the definition of quality as fitness for purpose (i.e., or use). This is a concept of quality defined by Juran and Godfrey (1999), and according to OECD (2008, p.44), “this definition is broader than has been used in the past when quality was equated with accuracy”, even more, “The most important quality characteristics depend on user perspectives, needs and priorities, which vary across user-groups”.

The San Francisco Declaration on Research Assessment (DORA) is a set of guidelines directed to funding agencies, institutions, publishers, researchers, and organizations that supply metrics (AMERICAN SOCIETY FOR CELL BIOLOGY, 2015). Particularly, the DORA asks about the use of the Journal Impact Factor researchers' quality assessment (OECD, 2016).

The Metric Tide is an independent review on the role of metrics in research assessment and management, produced by a group of scientometricians and bibliometricians that analysed the 2014 UK Research Excellence Framework (REF2014) (WILSDON et al., 2015). The review identified 20 specific recommendations to the next cycle of REF at the United Kingdom.

The Leiden Manifesto (HICKS et al., 2015) is a general set of principles for research metrics. In response to criticisms (see Frame 4 ), a group of researchers gathered at the 2014 International Conference on Science and Technology Indicators (STI2014) to produce a set of 10 principles that include attention to transparency, flexibility, and context, among others. Figure 6 lists the 10 principles of the Leiden Manifesto.

Figure 6 – The 10 principles of the Leiden Manifesto for research metrics



Source: Adapted of Hicks et al. (2015)

The 10 principles of the Leiden Manifesto (HICKS et al., 2015) start by recommending the use of objective metrics, and that they are used in support of qualitative judgement. To contextualize its purpose, similarly to the definition of quality as fitness for purpose (OECD, 2008), they suggest aligning the metrics with the mission and purposes of institutions. They emphasize the need to acknowledge local instead of universal research, and local rather than universal citation practices. In agreement with Lane (2010), they suggest that measures should be transparent, so data and analyses can be verifiable. They stress the importance of reviewing a researcher's trajectory, promoting comprehensive assessments. They warn about the risk of unfair results when specifying information at different levels of abstraction. They also warn about researchers directing their work based on the metric and not based on broader impacts. They conclude emphasizing the importance of data quality and recency.

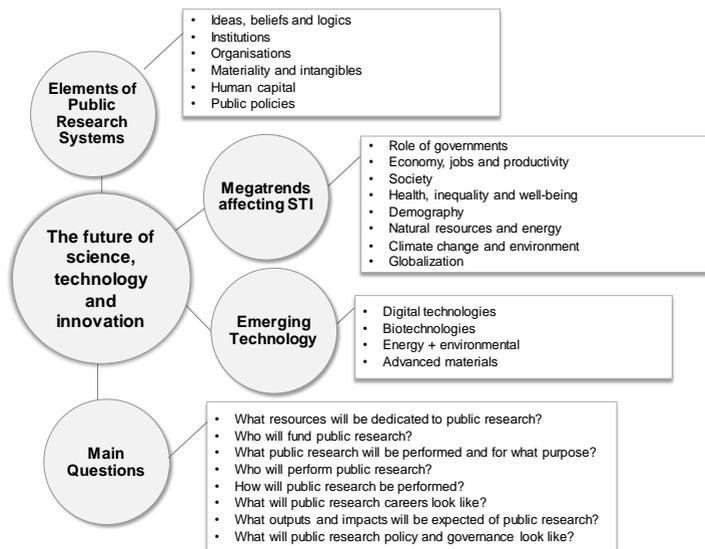
The Leiden Manifesto has motivated a series of studies since the STI2014 conference, they have examined each one of its 10 principles, and discussed its overall relevance and usefulness (BORNMAN; HAUNSCHILD, 2016; COOMBS; PETERS, 2017; DAVID; FRANGOPOL, 2015; GADD, 2015; MARZOLLA, 2016; WILDGAARD et al., 2016). In general, researchers have agreed with the relevance of the Leiden Manifesto and recommended its application on bibliometrics and related fields. I will list three special examples: One is the study of Bornmann and Haunschild (2016) that investigated the use

of the Leiden Manifesto in Altmetrics, and concluded that the principles are of great importance and should be taken into account. The other is a critical discussion of the Leiden Manifesto for libraries, which was conducted by the Association of European Research Libraries (LIBER). The final recommendation is that all libraries should embrace the Manifesto's principles (COOMBS; PETERS, 2017). The third is that the Leiden Manifesto was recently awarded with the 2016 EASST Ziman award for a “*collaborative promotion of public interaction with science and technology*” (LEIDEN MANIFESTO BLOG, 2016).

### 2.2.7 Future trends on research assessment.

The OECD (2016) attempts to see the future of Science, Technology and Innovation (STI) in a time horizon of 10-20 years. The elements of public research systems, megatrends affecting STI, and emerging technologies, are analyzed to answer eight key questions that impact public research policy. Based on this study, I seek to identify potential trends on research assessment in each one of these eight questions. An overview of the future of STI given by the OECD (2016) is illustrated in Figure 7.

Figure 7 – The overview of the future of STI



Source: Adapted of OECD (2016)

According to OECD (2016), the global capacity of Research and Development (R&D) has doubled in the last 15 years, and is expected to keep growing. However, the resources dedicated to research, which used to be funded by governments will shift. They tend to be increasingly provided by public-private partnerships, which will help to mobilize new sources of funding for the development of megatrends (e.g., health, energy, and globalization).

Other aspect observed in the OECD study is that public research has progressively shifted towards universities, by linking teaching and research through collaborative activities, such as, joint projects, PhD training, co-publication, joint appointments, among others. Furthermore, the involvement of citizens and organized groups in scientific efforts will contribute to develop a new culture of collaborative awareness. Another novelty is the “*do-it-yourself science*”, where citizens and organized groups conduct their own experiments.

Some emerging digital technologies, based mainly on Artificial Intelligence (AI) and big data analytics are modifying the way science is conducted and its results published, as pointed out by OECD (2016). With new technologies, research collaboration is becoming more capable, being able to promote cost sharing, academic mobility, and thus the expansion of research projects. In addition, new paradigms of research collaboration, such as, “*open science*”, “*open access*”, “*open data*”, and “*open collaboration*”, shift traditional bibliometric methods to alternative metrics, or Altmetrics. However, despite the strong use of technologies, the study considers that peer-review will remain an important mean for assessing research quality. Furthermore, the study also emphasizes the recommendations of Hicks et al. (2015) in the Leiden Manifesto for research metrics, which promotes the idea that research should be assessed on its own merits.

As a final point, the OECD (2016) poses the efforts of governments towards more agile STI policies. This effort includes regulations and governance arrangements; the creation of data infrastructures based on big data analytics for more evidence-based policy; the use of methodologies and indicators based on responsible metrics; and the encouragement of international connections for better data exchange, information, expertise and good practices.

## 2.2.8 Concluding remarks

Section 2.2 begins by posing research assessment, which according to Moed (2011) is “*research assessment includes the evaluation of research quality and measurements of research inputs, outputs and impacts, and embraces both qualitative and quantitative methodologies*”. After that, the main terms of this citation, which are research, quality, assessment, and measurement, were investigated. The rest of this section explored research assessment from its historical context to its future trends. Frame 5 summarizes section 2.2 in keywords that represent the subjects investigated in each subsection.

Frame 5 – Summary of the section 2.2

Research Assessment	Subsection	Keyword
	Terms and concepts	Science      Scientist Research      Quality Assessment      Evaluation Measure      Performance Metrics Estimation
	Historical context	Positivism      Scientific Method Governance      Research policy Internet      Electronic publication
	Current methods	Peer-review Bibliometrics      Scientometrics Informetrics      Altmetrics
	Performance metrics	SCI      Citation count IF      JIF      AIF Web-based metrics
	Critical aspects	Purpose      Context Data quality      Transparency General use      Normalization
	Source of principles for assessing research quality	Fitness for use DORA      Metric TIDE Leiden Manifesto
	Future trends	Research systems      Megatrends Data Science      R&D Altmetrics Leiden Manifesto Collaboration

Source: The author, 2017.

The terms and concepts subsection shows the root of terms associated to research assessment found on dictionaries.

The historical context subsection can be divided in three periods: *The first period* is from 16<sup>th</sup> to 19<sup>th</sup> centuries, which include the genesis of the Positivist movement, of the scientific method, and of the terms science and scientist.

*The second period* is the 20<sup>th</sup> century, especially after 1960s, which is considered the era of quantitative computer-tabulated science metrics. In the 1970s and 1980s, practices of research governance motivate bibliometric studies. In 1990s, the new age of connectivity, supported by the Internet and the Web, created an adequate environment to collaborative practices on research assessment.

*The third period* is the 21<sup>st</sup> century, in which the scientific community has witnessed the growth of electronic publications, and consequently, research policy decisions increasingly based on quantitative indicators.

The current methods subsection identifies five essential fields of study on research assessment: Peer-review is the oldest and most common field, it is based on a subjective process of research evaluation (SPIER, 2002); Bibliometrics is "*the application of mathematics and statistical methods to books and other media of communication*" (PRITCHARD, 1969, p. 349); Scientometrics involves "*the quantitative methods of the research on the development of science as an informational process*" (NALIMOV, 1971, p. 2); Informetrics encompasses "*The study of the application of mathematical methods to the objects of information science*" (NACKE, 1979, p. 220); Altmetrics, or Scientometrics 2.0, *gathers a set of web-based metrics associated to social media environment* (PRIEM, 2010).

The performance metrics subsection investigates metrics based on current methods (i.e., Bibliometrics, Scientometrics, Informetrics, and Altmetrics). These metrics are derived from the Science Citation Index (SCI) that was created by Eugene Garfield in 1960s, and since then many other metrics have been developed (VAN NOORDEN, 2010).

The critical aspects subsection lists criticisms concerning the misapplication of research metrics. These aspects are categorized in seven items: general use, purpose, context, data quality, time in metrics, normalization, and transparency and flexibility.

The source of principles for assessing research quality subsection presents recommendations for the appropriated use of performance metrics on research assessment. This review found four studies on this matter, which are: The Handbook on Constructing Composite Indicators

(OECD, 2008), The San Francisco Declaration on Research Assessment (American Society for Cell Biology, 2012), The Metric Tide (WILSDON et al., 2015), and The Leiden Manifesto for Research Metrics (HICKS et al., 2015). Among these four studies, the Leiden Manifesto is the most suitable set of bibliometric recommendations for research assessment in general, which gathers, in only 10 principles, many of the ideas proposed by the three other studies.

The relevance of the Manifesto (HICKS et al., 2015) is evidenced by several studies (BORNMANN; HAUNSCHILD, 2016; COOMBS; PETERS, 2017; DAVID; FRANGOPOL, 2015; GADD, 2015; MARZOLLA, 2016; WILDGAARD et al., 2016). I will list three significant examples: Bornmann and Haunschild (2016) suggested the application of the Leiden Manifesto principles in Altmetrics. Coombs and Peters (2017) described the study of the Association of European Research Libraries (LIBER), which recommended libraries to use the Manifesto's principles. At last, the Leiden Manifesto was recently awarded with the 2016 EASST Ziman award for a '*collaborative promotion of public interaction with science and technology*' (LEIDEN MANIFESTO BLOG, 2016).

The future trends subsection is based on the study of the OECD (2016) about the future of science systems over a time horizon of 10-20 years. Elements of public research systems, megatrends affecting STI, and emerging technologies are analyzed to answer eight key questions that impact the public research policy (see Figure 7).

In summary, this review identified five key future trends that directly impact research assessment: First, the global capacity on R&D will continue to grow, demanding agile science and technology policies. Second, public research will progressively shift towards universities by linking teaching and research through collaborative activities. Third, the establishment of a new emerging culture on STI, "*do-it-yourself science*", that involves citizens and organized groups conducting their own experiments; Fourth, the expansion of digital technologies based mainly on artificial intelligence (AI) and big data analytics, by supporting new paradigms such as, "*open science*", "*open access*", "*open data*", and "*open collaboration*", which among others, shift traditional Bibliometrics to Altmetrics. Fifth, OECD (2016) promotes the idea that research should be assessed on its own merits, by emphasizing the recommendations of Hicks et al. (2015) in the Leiden Manifesto for Research Metrics.

## 2.3 RESEARCH COLLABORATION

In this section, by looking at the research collaboration as a construct inserted in the context of the research assessment, I review the literature under this new focus, that is, seeking for the same elements previously studied but in the light of research collaboration. To achieve this goal, this review is outlined in seven subsections, as described in the next paragraphs.

In the first subsection, this investigation looks back at the 17th and 18th centuries to show early initiatives of the working together of researchers, and its connection to the origins of modern science.

The second subsection provides a brief view of the attempts to explain the theoretical bases of research collaboration.

In the third subsection, three terms and concepts associated to research collaboration are examined in an etymological perspective, which are, collaboration, cooperation, and co-production.

The fourth subsection describes the most common methods and performance metrics applied on research collaboration. Such methods and metrics are based on bibliometric analysis, and it has origin on metrics already cited in section 2 – Research Assessment.

The fifth subsection investigates sources of principles for assessing research quality, such as studied on section 2.2.6, however, with focus on research collaboration.

The sixth subsection searches in literature for related projects in research collaboration, in order to identify their difficulties, solutions, as well as, their contributions to the conduction of big science projects.

At the end, the seventh subsection concludes this review and focus on future trends on research collaboration assessment, based on the perspective of OECD (2016).

### 2.3.1 The historical context of research collaboration

The early efforts toward an exchange of knowledge between scientists were promoted by the Royal Society, and started in the 17<sup>th</sup> century (BEAVER; ROSEN, 1978; PETERS, 2006). Examples of these initiatives are regular meetings, exchange of letters, and travels of scientists to meet other European scientists. Other evidence of collaborative activities in this early years of modern science is the first collaborative paper published in 1665 (BEAVER; ROSEN, 1978).

A second period of growth in research collaboration was observed by Peters (2006), which he called “*colonial science*” referring

to the colonial expansion through the great navigations in the 19<sup>th</sup> century. In this period, the scientific community had become more professional through the foundation of scientific societies, in which the collaboration between scientists was used as a mechanism of recognition, visibility and productivity of researchers (BEAVER; ROSEN, 1978).

However, it was only in the 20<sup>th</sup> century that research collaboration was effectively established (around the 1950s) as result of the emergence of “*Big Science*”, which was evidenced by the studies of Derek de Solla Price (PETERS, 2006), in his book “*Little Science, Big Science*” (PRICE, 1963). In this study, the author analyzed research production after World War II through the SCI (GARFIELD, 1955, 1964), and noticed that research collaboration was becoming a dominant mode of research (FIORE, 2008; YAGI; BADASH; DE BEAVER, 1996).

In this time, studies on research collaboration based on publications analysis proliferated, which contributed to building the structure of research collaboration that is currently known. For example, other two significant works by Price were “*Networks of scientific papers*” (PRICE, 1965) and “*Collaboration in Invisible College*” (PRICE; BEAVER, 1966). In the first, Price and Donald Beaver worked together and introduced studies on connectivity of scientific papers, co-citation networks, and mapping of scientific fields, resulting on the first techniques for research collaboration assessment (LEYDESDORFF, 1998; YAGI; BADASH; DE BEAVER, 1996).

These pioneer contributions had strong influence over new generations of authors, such as Henry Small, who gathered on the study “*Co-citation in the scientific literature: A new measure of the relationship between two documents*” (SMALL, 1973), a set of techniques on co-citation. Examples of such techniques are: clusters of co-citations, citation maps, and patterns of scientific collaboration, which demonstrated direct or indirectly, the cooperativity among researchers, and the interdisciplinarity of different fields (SMALL, 1973, 1999).

Around the end of the 1970s, the “*Studies in Scientific Collaboration - Part I, II and III*” (BEAVER; ROSEN, 1978, 1979a, 1979b) investigated the root of co-authorship, and formally acknowledged research collaboration as co-authorships of scientific papers, incorporating the terms “*collaboration*”, “*teamwork*”, “*mutual, cooperative, or joint research*”, and “*joint or co-authorship*”, as synonyms of research collaboration.

In addition to studies based on co-citations, a piece of work emphasizing the primary authors as unit of analysis was presented by White and Griffith (1981). This study was based on metrics to construct

and analyze “*map of authors*”. These maps are used to identify author groups, proximities between authors, and clusters of authors.

“*What is research collaboration?*” (KATZ; MARTIN, 1997) is another significant study, which addresses the classical concept of research collaboration, defined as “*the working together of researchers to achieve the common goal of producing new scientific knowledge*” (1997, p. 7).

In the last decades, research collaboration has become crucial for productivity in science and engineering (e.g., ADAMS, 2014; BOZEMAN; FAY; SLADE, 2013; KUMAR, 2015; LEAHEY, 2016). In the next sections, the current context and trends of research collaboration will be included with more details.

### **2.3.2 The theoretical bases of research collaboration**

As said by Shrum et al. (2007, p. 7) “*to date, no comprehensive theory of scientific collaboration exists*”. Despite Shrum’s statement, this review scrutinizes the literature searching for the theoretical bases of “*research or scientific collaboration*”, and presents its main findings in the following paragraphs.

The first study to propose a theory on scientific collaboration was Beaver and Rosen (1978), which was followed by two other studies of Beaver and Rosen (1979a, 1979b) that investigated the root of scientific collaboration from the 17<sup>th</sup> to 20<sup>th</sup> centuries. This trilogy has a historical and sociological perspective, as previously mentioned on section 2.3.1., and suggests that “*scientific collaboration is a response of the professionalization of science*” (BEAVER; ROSEN, 1978, p. 64).

In addition to the perspective of professionalization of science, Beaver and Rosen (1978) also suggest there is an association between research collaboration and the specialization of science, which was originated in the 19<sup>th</sup> century and outlined the fields of science (e.g., Physics, Mathematics, Biology, etc). The explanation for this association is that collaboration between researchers encourages scientists to cross bounds of scientific fields (BEAVER; ROSEN, 1978, 1979a; BELLOTTI; KRONEGGER; GUADALUPI, 2016; LEAHEY, 2016; LEAHEY; REIKOWSKY, 2008; MOODY, 2004).

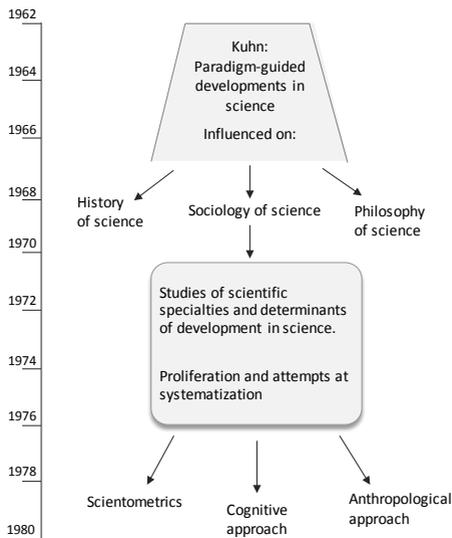
More recent studies, in the same historical and sociological vein, consider research collaboration as an emergent phenomenon of global science (HENNEMANN; RYBSKI; LIEFNER, 2012; LEAHEY, 2016; PETERS, 2006). According to this perspective, despite the fact that research collaboration as a sociological phenomenon had originated in the

17<sup>th</sup> century during the scientific revolution, it was only effectively established in the 20<sup>th</sup> century (PETERS, 2006). In fact, as observed by Price (1963) in his book “*Little Science, Big Science*”, scientific research shifted in the early 1960s from the “*solo investigators to team scientists*” (LEAHEY, 2016, p. 81).

After the second war, studies on social science (KUHN, 1962), Social Constructionism (BERGER; LUCKMANN, 1966), Artificial Intelligence (MINSKY, 1961), Cybernetics (WIENER, 1961), Cognitive Science (MILLER, 2003), and Bibliometrics and Scientometrics (GARFIELD, 1963; HAUSTEIN; LARIVIÈRE, 2015; PRICE, 1963) brought, directly or indirectly, significant contributions to the theoretical bases of research collaboration.

The influence of Thomas Kuhn’s study on social science is synthesized by Rip (1981) in a chronological model, as illustrated in the Figure 8. In this model, the author emphasizes the importance of looking at historical, social and cognitive aspects of evolution of scientific specialties and disciplines, and its attempts at systematization through scientometrics, cognitive and anthropological approaches.

Figure 8 – The study of dynamics of science from 1962 to 1980, according to Kuhn (1962).



Source: Adapted of Rip (1981, p.297).

The study of Kuhn (1962) does not address research collaboration specifically, however, its definition of scientific community was widely used to conceptualize research collaboration (e.g., KATZ; MARTIN, 1997). Thomas Kuhn defined scientific community as a set of scientists that share the same paradigm (i.e., a set of beliefs, values, techniques and models). This community is characterized by the practice of a scientific specialty, its members have similar backgrounds, and pursue a set of shared goals in a common professional judgment.

Another example of the sociological influence on research collaboration is the cognitive science (CS), “the product of a time when psychology, anthropology and linguistics were redefining themselves and computer science and neuroscience as disciplines were coming into existence” (MILLER, 2003, p. 141). Cognitive science is a combination of at least six interdisciplinary fields of inquiry, psychology, linguistics, neuroscience, computer science, anthropology and philosophy, in an attempt to understand brain, cognition and behavior (DICKINS, 2004; MILLER, 2003).

Currently, multiple lines of investigation following a historical and sociological perspective are found in the literature. *Interdisciplinarity*, *science of team science*, *cyberinfrastructure*, *e-science*, and meta-knowledge are some examples of these lines.

*Interdisciplinarity* (KLEIN, 2008) has as one of its motivations, the notion that future scientific advances, which can significantly influence human life, will come from collaborations of researchers from multiple disciplines.

*The science of team science* (STOKOLS et al., 2008) is an interdisciplinary practice in medical fields, particularly in clinical applications.

*Cyberinfrastructure* (ATKINS, 2003) and *e-science* (HEY; HEY, 2006) are two similar concepts concerned with virtual infrastructures for scientific collaborations.

*Meta-knowledge network* (EVANS; FOSTER, 2011) is a new potential field of study, which seeks to understand the entire cycle of knowledge, and to create knowledge about knowledge.

### **2.3.3 Distinct meanings for the term “collaboration”**

The goal of this subsection is to investigate the concept of collaboration, its different combinations and related terms, to better delimitation of the scope of this study.

Etymologically, the word “*collaboration*” is the combination of two words “*col*” that means “*together*” and “*labor*” that means “*work*” (BORROR, 1960). For example, Mauthner and Doucet (2008, p. 974) point out that collaboration is related to “*labor to be shared*”. Another meaning of collaboration concerns “*co-elaboration*”, for instance, “*Individuals co-elaborate a shared understanding of the problem*” (DÉTIENNE; BAKER; BURKHARDT, 2012, p. 2). In addition, collaboration is influenced by aspects such as, reciprocity, willingness of parties, risk sharing, responsibilities, and high levels of trust (POOCHAOREN; TING, 2014).

A related term to collaboration is “*cooperation*”. Borrer (1960) defines cooperation, etymologically, as the combination of “*co*” that means “*together*” and “*opera*” that means “*work*”. Hence, cooperation is a synonym of collaboration. However, Roschelle and Teasley (1995, p. 70) distinguish the terms by considering cooperation “*an activity where each person is responsible for a portion of the problem solving*”, which differs from collaboration where all members of the group share the understanding of the problem before executing a task. For example, each participant of a group receives a specific task to be carried out independently, under the instructions of a coordinator (DÉTIENNE; BAKER; BURKHARDT, 2012).

“*Coproduction*” is another term also associated to collaboration (POOCHAOREN; TING, 2014). Coproduction etymologically originated from “*co*” that means “*together*” and “*prod*” that means “*reveal*” (BORROR, 1960). In a general concept, coproduction predicts the idea of citizen participation in the provision of any service (BRUDNEY; ENGLAND, 1983). In this vein, the classical concept of coproduction was defined by Ostrom (1996, p. 1076) as “*the process through which inputs used to produce a good or service are contributed by individuals who are not “in” the same organization*”.

In the research context, Mobjörk, (2010) distinguishes collaboration in three cross-disciplinary approaches: *Multidisciplinary Collaboration*, the cooperation between researchers within each discipline, and that can be understood as “*a division of labour*”; *Interdisciplinary Collaboration*, which is characterized by collaboration between researchers from different disciplines, with a common methodological approach and a shared problem formulation; and *Transdisciplinary Collaboration*, which is based on the classical concept of Klein (2004, p.517), in which “*Transdisciplinarity moves beyond “interdisciplinary” combinations of academic disciplines to a new understanding of the relationship of science and society*”.

### 2.3.4 Current methods and performance metrics applied on research collaboration

In section 2.3.1, I cited some of the first techniques used to measure research collaboration, which were based on publication analysis, and derived from the Science Citation Index (SCI) (GARFIELD, 1964). Examples of such early techniques are synthesized in Frame 6 (LEYDESDORFF, 1998; PRICE, 1965; SMALL, 1973; WHITE; GRIFFITH, 1981; YAGI; BADASH; DE BEAVER, 1996).

Frame 6 – The early techniques to measure research collaboration

<b>Metric</b>	<b>Description</b>
Citation analysis (GARFIELD, 1964)	<i>“In citation analysis, the emphasis is on the number of citations received by an article, journal or author, or which article cites which article”</i> (KUMAR, 2015).
Co-citation of papers (SMALL, 1973)	<i>“Co-citation is defined as the frequency with which two documents are cited together”</i> (SMALL, 1973, p. 28)
Author co-citation analysis (ACA) (WHITE; GRIFFITH, 1981)	<i>“Co-citation of authors results when someone cites any work by any author along with any work by any other author in a new document of his own”</i> (WHITE; GRIFFITH, 1981, p. 163).

Source: The author, 2016.

Despite studies on citation analysis only emphasizing networks of scientific papers, according to Yagi, Badash and Beaver (1996), such studies motivated Derek de Solla Price and Donald Beaver to map research groups, which lead them to propose the idea of co-authorship of scientific papers. Co-authorship consists of crediting contributions of each co-author in two different ways. One is the full method, in which each author receives full credit for the paper. Another is the fractional method, in which each one of the authors receives a fraction of the credit. This method is still widely used to measure research collaboration nowadays (BOZEMAN; FAY; SLADE, 2013; KUMAR; JAN, 2013), it incorporates the concept of research cooperation between co-authors (VINKLER, 1993), and several indicators have been developed influenced by its principle, as related in Frame 7.

Frame 7 – Indicators and research collaboration methods based on the co-authorship principle

Metric	Description
Degree of collaboration (DC) (SUBRAMANYAM, 1983)	In a set of papers, this index is represented by the number of authors divided by the total number of papers.
Collaborative Index (CI) (LAWANI; ROAD, 1986)	This index describes the average number of authors per paper for a given set of papers.
Collaborative Coefficient (CC) (AJIFERUKE; BURELL; TAGUE, 1988)	This coefficient is a combination of both DC and CI from a field in a single value.
Productivity (P) (ABRAMO; D'ANGELO, 2011, p. 353)	The Productivity (P) is the total of publications authored by a scientist during an observed period.
Fractional Productivity (FP) (ABRAMO; D'ANGELO, 2011, p. 353)	The FP is the “ <i>total of the contributions to publications authored by a scientist, with “contribution” defined as the reciprocal of the number of co-authors of each publication</i> ”.
Fractional Scientific Strength (FSS) (ABRAMO; D'ANGELO, 2016, p.598)	$FSS = \frac{1}{t} \sum_{i=1}^N \frac{C_i}{\bar{C}} f_i$ , where: <p><i>t</i> is the number of years worked by a researcher in a period under observation;</p> <p><i>N</i> is the number of researcher publications in a period under observation;</p> <p><i>C<sub>i</sub></i> is the number of citations received by publication <i>i</i>;</p> <p><math>\bar{C}</math> is the average of citation distributions received for all cited publications, in the same year and subject category of publication <i>i</i>;</p> <p><i>f<sub>i</sub></i> is the fractional contribution of a researcher to publication <i>i</i>.</p>
Co-authorship networks	It is an indicator of social connection, in which a set of algorithms based on social network analysis (SNA) and co-authorship is applied to better understand the topology of research collaboration networks (NEWMAN, 2001, 2004).

Source: The author, 2017.

In order to propose a set of “*statistical properties of co-authorship networks*” (see Frame 8) in which scientists are connected to each other by coauthored papers, Newman (2001) investigated the bases of *SNA*, as well as, co-authorship (PRICE, 1965), co-citation of papers (SMALL, 1973), and author co-citation analysis (WHITE; GRIFFITH, 1981).

*Social network analysis (SNA)* is an approach based on social science and graph theory that has been widely adopted for analysis and visualization of research collaboration networks (KUMAR, 2015; TSAI; LIAO, 2009; WASSERMAN; FAUST, 1994).

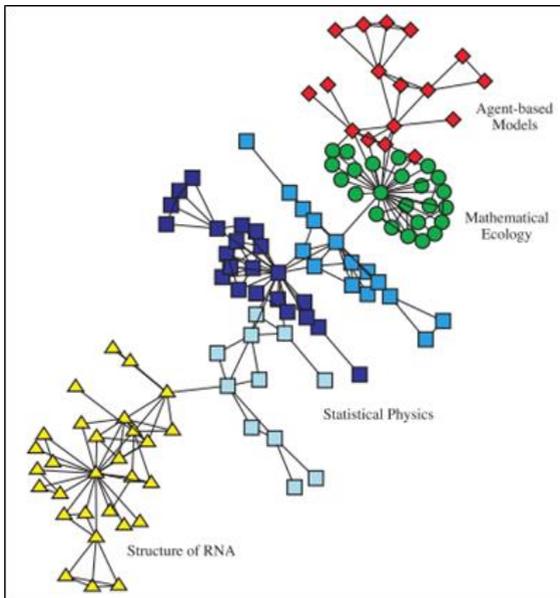
Frame 8 – Statistical properties of co-authorship networks based on Newman’s studies

<b>Statistical property</b>	<b>Description</b>
Number of authors	The total number of distinct authors found in the data sample.
Number of papers	The total number of papers found in the data sample.
Papers per author	The average number of papers published by an author.
Authors per paper	The average number of coauthors on a paper.
Average collaborators	The average number of collaborators of an author.
The giant component	The largest connected group of individuals in the network.
Average distance	The average vertex–vertex distance between connected individuals in the network
Largest distance	The largest distance between connected individuals in a network (i.e., diameter of the network).
Average degrees of separation	The average number of links in the network, between all pairs of scientists.
Clustering coefficient	The clustering coefficient, which is the probability that two coauthors will also be coauthors of one another.
Assortativity	The number of collaborators of adjacent vertices in the network. A positive value indicates that people tend to collaborate with others who have many collaborators.

Source: Adapted of Newman (2004, 2001)

The *co-authorship networks* emerged as an alternative to research collaboration assessment. It tries to identify relationship patterns between individuals, through the analysis of nodes (i.e., individual actors within the network) and ties (i.e., the relationship between the actors) in the network (KUMAR, 2015; ZHANG et al., 2013). The method was illustrated by Newman (2004), who used bibliographic databases in biology, physics, and mathematics, applied to a set of statistical properties, based on the SNA, to identify patterns of scientific collaboration. Figure 9 shows a small co-authorship network, in which the nodes represent authors, and the lines between two authors represent coauthored papers.

Figure 9 – An example of a small co-authorship network



Source: Adapted of Newman (2004, p.5206)

The method depicts several aspects of the co-authorship network, such as how fragmented or cohesive the community of authors is, who the key authors are, and how connected the best authors are. The method specially emphasizes the entire set of authors and their relationships, in contrast to the analysis of individual authors (KUMAR, 2015).

### 2.3.5 Source of principles for assessing research quality on collaboration

In subsection 2.2.6, four studies were presented four sources of principles for assessing research quality: The Handbook on Constructing Composite Indicators (OECD, 2008), The San Francisco Declaration on Research Assessment (American Society for Cell Biology, 2012), The Metric Tide (WILSDON et al., 2015), and The Leiden Manifesto for Research Metrics (HICKS et al., 2015). These studies attempt to give guidance on how to lead with the criticism concerning the misapplication of research metrics in general (see subsection 2.3.4). This review explored these four studies and no explicit reference to research collaboration was found.

Despite the fact there is not explicit criticism or recommendation elaborated for research collaboration, two aspects must be considered: One is the inappropriate application of research metrics. For example, David and Frangopol (2015, p.2256) point out that “*indicators designed to evaluate journals are wrongly used to evaluate individuals and/or groups*”. These authors comment that this issue is addressed by the Leiden Manifesto (HICKS et al., 2015). The other aspect concerns the second principle of the Manifesto, which addresses the purpose of the assessment. The second principle states that “*performance should be measured against the research goals of the institution, group or researcher*” (HICKS et al., 2015, p.430).

### 2.3.6 Related interdisciplinary collaborative projects

This subsection intends exemplify relevant initiatives on research collaboration, showing their challenges in the conduction of the collaborative project, and the approaches adopted by them. Thus, three examples of interdisciplinary collaboration are briefly described, which are two research projects, and one research group. According to Maglaughlin and Sonnenwald (2005, p.1), “*Interdisciplinary collaboration occurs when people with different educational and research backgrounds bring complementary skills to bear on a problem or task*”.

The *Human Genome Project* (COLLINS; MORGAN; PATRINOS, 2003), and the *CMS experiment* (CMS COLLABORATION, 2008) are two collaborative projects which have been categorized as “*grand challenges and great opportunities in science, technology, and public policy*” (OMENN, 2006). The third, the

*Microcephaly Epidemic Research Group* (MERG, 2016) is a research group, which originated in Brazil. These three examples are characterized by a type of research collaboration that Price (1963) called “*Big Science*”. Nowadays *Big Science* concerns *collaborative projects* that involve multiple investigators and conceptualizations of research problems from different institutions and cultures (e.g., COLLINS; MORGAN; PATRINOS, 2003; HILL et al., 2014; ORTOLL et al., 2014; WELSH; JIROTKA; GAVAGHAN, 2006).

#### 2.3.6.1 *The Human Genome Project (HGP)*

According to Collins, Morgan and Patrinos (2003), the human genome sequencing started in 1988 as an initiative of the government of the USA through the National Health Institute (NIH). This grand challenge project involved not only the NIH, but the collaboration of the Department of Energy (DOE), American private institutes, and the experience of renowned scientists. They represented universities, institutes and laboratories from many countries and fields, around a common goal: mapping and sequencing the human genome. One of the great initiatives of the team was sharing the knowledge produced, making it free and accessible to all. The HGP was declared complete in 2003, providing the scientific community with a legacy of contributions.

In the context of this thesis, the HGP is a reference on conducting a large collaborative project, in special the recruitment of the research team. Collins, Morgan and Patrinos (2003) relate that initially strategies adopted by the Manhattan Project (a top-down strategy) and the Hubble Space Telescope Project (a bottom-up strategy) were investigated. After that, a familiar peer-review process of funding biomedical research was adapted for HGP. This process was used for both, the recruitment and evaluation of researchers.

#### 2.3.6.2 *The CMS Experiment*

The Compact Muon Solenoid (CMS) detector operates at the Large Hadron Collider (LHC) at the CERN laboratory, near Geneva, Switzerland (CMS COLLABORATION, 2008). It is part of a large-scale collaborative project with another detector called the ATLAS experiment, and involves more than 5,000 researchers from dozens of institutions and countries, with the mission of studying the Higgs boson mechanism (CASTELVECCHI, 2015).

Despite the fact that the CMS experiment is a “*grand challenge*” project (OMENN, 2006), what is special for this thesis is the extreme number of authors registered in its publications, which emphasizes the collaboration between research teams. This high figure of authors has taken the attention of editors (e.g., CASTELVECCHI, 2015) and bibliometricians (e.g., WILSDON et al., 2015). For example, one of its papers published in 2008 became the first publication to have 3,000 authors, and another recent paper has 5,154 authors (CASTELVECCHI, 2015).

### 2.3.6.3 *The Microcephaly Epidemic Research Group (MERG)*

The Brazilian Microcephaly Epidemic Research Group (MERG) was created in 2015 during a public-health emergency of Microcephaly in the state Pernambuco. The goal of MERG was facing this public health crisis to better understand and attempt to solve it (BUTLER, 2016). The research group is registered at the Directory of Research Groups (DGP). It is based at the Aggeu Magalhães Research Center (CPqAM), which is a facility of the Oswaldo Cruz Foundation (Fiocruz), in Recife, Pernambuco. Furthermore, the MERG consists of researchers specialized in the fields of health and biology, from several institutions from Brazil, the UK, and the USA (DGP.CNPQ, 2015).

The researchers of MERG discovered a link between microcephaly and the Zika virus, and hence that the Zika virus causes microcephaly (GROUP, 2016). For leading this “*grand challenge*”, its leader was acknowledged as one of the most renowned scientists in the world (BUTLER, 2016), because as stated by Frieden (2017), “*She understood that this was a global crisis requiring global collaboration*”.

### 2.3.7 **Future trends on research collaboration**

This subsection extends subsection 2.2.7 – Future trends on research assessment, which are based on OECD (2016), and highlights three main trends: The international collaboration, the open collaboration, and the *do-it-yourself science*.

From the perspective of international collaboration, the “*grand challenges*” of science call for more cooperation and collaboration in STI policy. For example, due the large amounts of public research funding required, collaboration between institutions and researchers has been motivated in the field of research infrastructures.

OECD (2016) also takes into consideration that digital technologies have boosted radical changes in the way science is conducted. This fact is steering science toward the "*open science*" paradigms (OECD, 2015a). For example, "*open collaboration*", which is a capacity of online environments that "*offer new opportunities for people to form ties with others and create things together*" (FORTE; LAMPE, 2013, p.536).

Finally, the third trend is the expression "*do-it-yourself science*", which refers to a new culture of collaborative awareness, in that, citizens and organized research groups conduct their own experiments in co-production.

### 2.3.8 Concluding remarks

In this subsection, the intention is to synthesize the main findings of this literature review on research collaboration. Frame 9 summarizes section 2.3 in keywords that represent the subjects investigated in each subsection.

The *historical context* subsection described facts that evidenced collaborative practices between researchers through time. For instance, the first collaborative paper published in the 17th century; the foundation of scientific societies in the 19th century; and also, the emergence of "*Big Science*" in the 20<sup>th</sup> century. After the second World War, research collaboration was effectively established, and authors such as Derek de Solla Price, and Donald Beaver introduced the first studies on co-authorship of scientific papers.

In the *theoretical bases of research collaboration*, despite "*to date, no comprehensive theory of scientific collaboration exists*" (SHRUM; GENUTH; CHOMPALOV, 2007, p. 7), there are some assumptions about this theme. The first attempt to propose a theory on collaboration was from Beaver and Rosen (1979a, 1979b), who in a historical and sociological perspective, suggested that scientific collaboration is a response to the professionalization of science. Furthermore, these authors also suggest an association between research collaboration and the specialization of science. Other authors, following the vein of professionalization of science, consider research collaboration as an emergent phenomenon of global science (e.g., HENNEMANN; RYBSKI; LIEFNER, 2012; LEAHEY, 2016; PETERS, 2006). In addition to this perspective, cognitive aspects are also taken into account, such as the studies on cognitive science, which is composed of at least six interdisciplinary fields: psychology, linguistics, neuroscience, computer

science, anthropology and philosophy. It is also important to highlight the contributions of Thomas Kuhn's studies on the theoretical bases of research collaboration, which was synthesized by Rip (1981) in a chronological model, as illustrated in the Figure 8.

Frame 9 – Summary of the section 2.3

Research Collaboration		Subsection	Keyword
Research Collaboration	Historical context	Early efforts	17 <sup>th</sup> century
		Second period	19 <sup>th</sup> century
		Third period	20 <sup>th</sup> century
	Theoretical bases	Professionalization of science	Specialization of Science
		Global Science	Cognitive science
	Distinct meanings for the term collaboration	Collaboration	Cooperation
		Coproductio	
Performance metrics	Citation analysis	Co-authorship	
		Co-authorship networks	
Source of principles for assessing research quality	<i>Fitness for purpose</i>	DORA	Metric TIDE
		Leiden Manifesto	
Related interdisciplinary collaborative projects	The Human Genome project	The CMS experiment	The MERG group
Future trends	International collaboration	<i>Open collaboration</i>	<i>do-it-yourself science</i>

Source: The author, 2017.

In the subsection, *Distinct meanings for the term "collaboration"*, the intention was to investigate the concept of collaboration, its different etymologic combinations and related terms, to better delimitation of the scope of this study. The analyses of the terms "*cooperation*", "*collaboration*" and "*coproduction*", resulted in a categorization of their characteristics in evolutive process. Frame 10 presents this categorization emphasizing the limits of each concept.

Frame 10 – Conceptual characteristics of cooperation, collaboration and coproduction

Characteristics	Concepts		
	Cooperation	Collaboration	Coproduction
Work together	x	x	x
Produce together	x	x	x
Reveal together	x	x	x
Work divided	x	x	x
Labor divided	x	x	x
Responsibility divided	x	x	x
Multidisciplinary	x	x	x
Work shared		x	x
Labor shared		x	x
Responsibility shared		x	x
Value of reciprocity		x	x
Willingness of parties		x	x
High level of trust		x	x
Common goal		x	x
Interdisciplinary		x	x
Engagement with society			x
Multidimensional			x
Transcultural			x
Transdisciplinary			x

Source: The author, 2016.

In the *current methods and performance metrics applied on research collaboration* subsection, it was found that co-authorship of scientific papers, proposed by the studies of Derek de Solla Price and Donald Beaver (YAGI; BADASH; DE BEAVER, 1996), is still the most widely used metric for research collaboration (BOZEMAN; FAY; SLADE, 2013; KUMAR; JAN, 2013). Co-authorship implies crediting contributions of each co-author in two different ways. One is the full method, in which each author receives full credit for the paper. Another is the fractional method, in which each one of the authors receive a fraction of the credit. Furthermore, from co-authorship many other metrics and indicators have been proposed, for example, fractional

productivity (FP) (ABRAMO; D'ANGELO, 2011), and co-authorship networks (NEWMAN, 2001, 2004). The last example, co-authorship networks, is based on social network analysis (SNA), it tries to identify relationship patterns between individuals, through the analysis of nodes and ties in the network. The method particularly emphasizes the entire set of authors and their relationships, in contrast to the analysis of individual authors (KUMAR, 2015).

In the *source of principles for assessing research quality on collaboration* subsection, this review examined the four studies previously mentioned on subsection 2.2 – Research Assessment, which are: The Handbook on Constructing Composite Indicators (OECD, 2008), The San Francisco Declaration on Research Assessment (American Society for Cell Biology, 2012), The Metric Tide (WILSDON et al., 2015), and The Leiden Manifesto for Research Metrics (HICKS et al., 2015). However, there is no explicit reference to the use of quality references for research collaboration. Because these studies are generic recommendations for research evaluation, their application on research collaboration is still a little vague, and the only recommendation found is the second principle of the Leiden Manifesto, which states that “*performance should be measure against the research goals of the institution, group or researcher*” (HICKS et al., 2015, p.430).

The *related research collaboration projects* subsection briefly describes three relevant examples of projects on research collaboration. The Human Genome Project (COLLINS; MORGAN; PATRINOS, 2003), the CMS experiment (CMS COLLABORATION, 2008), and the Microcephaly Epidemic Research Group (MERG, 2016). These three examples are characterized by “*Big Science*” (PRICE, 1963), and “*grand challenges of science*” (OMENN, 2006). The HGP project illustrates the problem of conducting large collaborative projects, specially the recruitment of the research team. The CMS experiment illustrates the increasing number of co-authors registered in a paper, for instance, a unique paper had 5,154 authors, and this fact has called the attention of bibliometricians (CASTELVECCHI, 2015). The Brazilian Microcephaly Epidemic Research Group (MERG) is a local example of “*grand challenges of science*”. This research group was created during a public-health emergency of Microcephaly in the state of Pernambuco, and as a result, the researchers of MERG discovered a link between microcephaly and the Zika virus (GROUP, 2016).

Finally, in the future trends on research collaboration subsection, this review stresses three main trends (i.e., the international collaboration, the open collaboration, and the do-it-yourself science), which are based

on OECD (2016), and can be used to predict a time horizon of 10-20 years. The trend in overall collaboration, is due to a new collaborative culture that promotes the development of joint projects, co-publications, public-private partnerships, the involvement of citizens and organized groups in science, and hence, the encouragement of international connections. The open collaboration is a result of the influence of digital technologies, and it is a characteristic of the "*open science*" paradigm. The expression do-it-yourself science refers to a new culture of collaborative awareness, in that, citizens and organized research groups conduct their own experiments in co-production.

## 2.4 SOURCES OF KNOWLEDGE FOR RESEARCH ASSESSMENT

A particular concept of knowledge is given by Davenport and Prusak (1998, p. 5) as "*a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information*". According to these authors, this definition expresses characteristics that emphasizes the value of knowledge. They add that knowledge is in the mind of people, and it is also embedded in documents, repositories, routines, processes, practices, and norms.

In a similar vein, Chiva (2005) explains that according to the *Cognitivist theory*, knowledge is considered as a commodity, located in people's minds, and in data (i.e., documents and databases). For the *Connectionist theory*, knowledge is generated through networks and relationships. Moreover, considering the *Auto-Poiesis theory* (VARELA; MATURANA; URIBE, 1974), knowledge is shared through communication, and its significance depends on context, point of view, and experience.

Considering the *Cognitivist theory*, in which knowledge is in the mind of people, that is, in the mind of scientists, collaborators and decision makers, and embedded in documents, and repositories (curriculum vitae, research publications, CV databases, and bibliographical databases). Also considering the *Connectionist theory*, for instance, the knowledge is in the *co-authorship networks*, and in bibliographical productions. Consequently, considering the *Auto-Poiesis theory*, in which the knowledge depends of experience of scientists, collaborators and decision makers, in this section the literature is reviewed in order to find sources of knowledge for research assessment. In the following subsections eight examples of such sources will be described.

### 2.4.1 Scientists and Researchers

A simple definition of the word “*researcher*” is given by the online dictionary (e.g., [www.vocabulary.com/dictionary/researcher](http://www.vocabulary.com/dictionary/researcher)) as “*a scientist who devotes himself or herself to doing research*”. Apart from “*scientist*”, the term “*researcher*” is also used as a synonym for: investigators, experimenters, analyzers, inquirers, and examiners (e.g., [www.thesaurus.com/browse/researcher](http://www.thesaurus.com/browse/researcher)). In other words, according to these two dictionaries, the word “*researcher*” is a generic term to denominate someone who conducts research applying a method.

More precisely, according to Snyder (2011), until the beginning of the nineteenth century, the “*men of science*” were known as “*natural philosophers*”. However, William Whewell (1794-1866) created the term “*scientist*” by analogy with “*artist*”, in 1833. He understood that modern science required a new terminology to express its advancements. At the time, scientists were researchers trained in science at the university, followed a scientific method, and would receive funding as member of some scientific society.

### 2.4.2 Research collaborators

Guided by the question “*Who are the collaborators?*”, Katz and Martin (1997, p. 2) answer that they could be any researchers who work together to advance scientific knowledge (KATZ; MARTIN, 1997). Taking into account this generic answer, the authors also suggest that research collaborators include: those who work together in research projects, or who make substantial contribution for them; co-authors in papers and research projects; and those responsible for a key step of a scientific research.

The term “*scientific collaborators*” is also adopted as synonym of “*research collaborators*” (e.g., CARO; CATALDI; SCHIFANELLA, 2012; JONKERS; CRUZ-CASTRO, 2013; KUMAR; JAN, 2013).

### 2.4.3 Research decision makers

“*Who is responsible for organizing and performing the evaluation?*” ask Molas-Gallart (2012). I reviewed the literature and found that research decision makers, research managers, research funders, research planners, research directors, research administrators, science policymakers, evaluators, evaluation committees, and researchers themselves, are examples of such people, who are responsible for

planning strategies to evaluate and conduct science (BOZEMAN; GAUGHAN; YOUTIE, 2014; LARGENT; LANE, 2012; THOMSON-REUTERS, 2008).

#### 2.4.4 Careers trajectories

According to Arthur et al.(1989, p.8), "*Career is the evolving sequence of a person's work experiences over time*". The terms career and trajectories are viewed as synonyms that describe the path followed by people during their work life (VALENDUC et al., 2009).

Over recent decades, studies on research career and trajectories have been addressed by different authors on three main aspects: One of them is productivity along the career trajectories, which reflects the evolution of researchers' accomplishments (LEE; BOZEMAN, 2005; PAN; FORTUNATO, 2014; UNGER; RUMRILL JR., 2013). Other aspect is mobility, which describes researchers' trajectories in multiple institutions and countries (BERNELA; MILARD, 2016; NOORDEN, 2012; SCELLATO; FRANZONI; STEPHAN, 2015). The third aspect concerns collaboration among researchers, which has been considered an important indicator of credibility and quality for career advancement (WOOLLEY; CAÑIBANO; TESCH, 2016).

#### 2.4.5 Curriculum vitae

The root of the term "*Curriculum vitae*" (e.g., [www.merriam-webster.com](http://www.merriam-webster.com)) comes from the word "*curriculum*" that means "*a course of study*". When associated with the word *vitae*, it means "*course of (one's) life*". Curriculum vitae is commonly abbreviated as "*CV*", and pluralized as both "*curricula vitae*" and "*curriculum*".

Particularly for academic researchers, the curriculum vitae is a historical document declared by themselves, which registers the researcher's education, professional positions, scientific accomplishments, research collaborations, and grant successes (CAÑIBANO; BOZEMAN, 2009; DIETZ et al., 2000; GAUGHAN; BOZEMAN, 2002).

The wealth of knowledge contained in curriculums make them an attractive and potential source for bibliometric studies on career trajectories (e.g., mobility and collaboration). They have provided indicators to evaluate policy decisions in universities, funding agencies, and governments (CAÑIBANO; BOZEMAN, 2009; CGEE, 2016; FURTADO et al., 2015; GAUGHAN, 2009; SANDSTRÖM, 2009). The

use of CVs facilitates studies on mobility (SANDSTRÖM, 2009). For example, they provide information on career shifts, such as, geographical (i.e., from one place to another) and positional (i.e., from one position to another). Similarly, CVs also reveal information on a broader set of research collaborative activities (CAÑIBANO; BOZEMAN, 2009).

Despite the advantages of adopting CVs as a data source, there are many methodological problems related to availability, coding, and quality CVs' data (DIETZ et al., 2000; SANDSTRÖM, 2009). For instance, Sandström (2009) explains that some CVs can be condensed or 'truncated', and with incomplete personal information. In addition, Dietz et al. (2000) highlight that because the information is self-declared, generally in semi-structured format, the quality of information is subject to valuable information loss or inclusion of non-relevant data.

In order to minimize the problems afore mentioned, some countries are adopting and motivating the use of digital databases of CVs (e.g. Brazil, Portugal, Spain, Norway), and this approach opens up new possibilities for CV data analysis (CAÑIBANO; BOZEMAN, 2009).

#### **2.4.6 Research publications**

Making the findings of research inquiry public is the original intention of scientists (e.g., BRUMBACK, 2012; MCDOUGALL-WATERS et al., 2015). Thus, books, articles, proceeding papers, and journal articles, are examples of this public scientific communication called research publications, or also, bibliographical production.

The first scientific journal, *The Philosophical Transactions of the Royal Society*, was published in 1665, and consisted of letters, reviews and summaries of books (MCDOUGALL-WATERS et al., 2015).

After 350 years of this first physical journal publication, the scientific community has witnessed a genuine revolution in the traditional medias for research publications. Due, for example, to the Web 2.0 services, their physical format have shifted to virtual formats (e.g., blog posts, interactive graphics and video) (PRIEM, 2010).

#### **2.4.7 Bibliographical databases**

The history of bibliographical databases started with Eugene Garfield in the 1950s, who developed the SCI's multidisciplinary database, and whose early goal was searching for scientists' papers, and their respective citations. Initially, the SCI database was printed, after that, it was distributed to experts on CD-ROM. Finally, Thomson Reuters

launched the Web of Science (WoS) platform in 2002, which incorporates the SCI's multidisciplinary database, making it widely accessible to all researchers (GARFIELD, 2007).

As an alternative to the Web of Science, SCOPUS was launched by the Elsevier publishing company in 2004 (HARZING; ALAKANGAS, 2016). The SCOPUS includes multiple scientific databases such as Elsevier, Springer, Wiley-Blackwell, Taylor and Francis, Sage, IEEE, and Emerald, among others (ELSEVIER, 2014). It includes references starting from 1966, and it has more than 20,800 peer-reviewed journals, and more than 5,000 international publishers in health, physical, social, and life sciences.

These two digital libraries, WoS and SCOPUS, cover a vast number of research productions (e.g., journals, papers, reviews, books, book chapters, editorials, etc.), and offer a set of tools to store and retrieve scientific literature from a long range of years (ELSEVIER, 2014; WALTMAN, 2016).

Google Scholar (GS) was presented by Google in 2004 as an open access alternative for scholarly Web databases in an attempt to democratize the access to research publications (WILSDON et al., 2015).

## 2.4.8 CV databases

According to Dietz et al. (2000), one motivation to study researchers is the ubiquitous utilization of *curriculum vitae* (CV). This advantage has been recently strengthened by the development of CV databases such as the Brazilian *Lattes* database which will be explored in the next section (e.g., CAÑIBANO; BOZEMAN, 2009; CNPQ, 2017; LANE, 2010; PACHECO et al., 2006).

### 2.4.8.1 *The Brazilian Lattes database*

Brazilian *Lattes* database was launched in 1999, by the National Council for Scientific and Technological Development (CNPq), as a tool to support activities and research policies of the Brazilian Minister of Science, Technology, Innovation and Communication (MCTIC).

The *Lattes* database (<http://lattes.cnpq.br/>) is the core of a government platform on science, technology and innovation (STI), called *Lattes* Platform. This platform, apart from the CV directory, gathers data

from the Institution directory (<http://di.cnpq.br/di/index.jsp>), and also from the Research Groups directory (<http://lattes.cnpq.br/web/dgp>).

According to the “*Painel Lattes statistics*” (<http://estatico.cnpq.br/painelLattes/>), there are 5,156,293 curriculums registered in the *Lattes* database, as of July 2017. The CVs are individually registered by students, professors, researchers, and practitioners from all research areas and institutions in Brazil. Each CV contains data on researcher identification, education, affiliations, funding, accomplishments and participations in conferences and committees (CNPQ, 2017).

Eighteen years after its creation, from 1999 to 2017, the *Lattes* database is now acknowledged as one of the most important information resources in STI, and some of its approaches have been examples of good practices. For instance, Cañibano and Bozeman (2009, p. 90) comment that the Lattes database allows “*one to freely access in the Internet a standardized and quite complete CV*”. But it was with the publication of Lane (2010), that the Lattes database became internationally perceived as “*one of the cleanest researcher database in existence*”, by providing “*high-quality of data*” (2010, p. 488). This accurate information, according to Perlin et al. (2017) is due to the fact that MCTIC requires researchers to keep their *Lattes* profiles updated, in order to participate in grants and evaluation processes. These authors checked the accuracy of the information for a sample of 180,000 CVs, and they found that 81.27 percent of the sample had been updated after 2014.

## 2.4.9 Concluding remarks

Section 2.4 first investigated what knowledge is, and found the classical definition of Knowledge of Davenport and Prusak (1998, p. 5), in which knowledge is stated as “*a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information*”.

In addition to the definition of knowledge, three theories about where knowledge is stored, were found in literature: the *cognitivist theory*, the *connectionist theory*, and the *auto-poiesis theory*. Chiva (2005) explains that for *cognitivists*, knowledge is located in data and people’s minds; and for the *connectionists*, knowledge is in networks and relationships. According to Varela et al. (1974), which proposes the *auto-poiesis theory*, knowledge is shared through communication, and its significance depends on context and experience.

Considering these three theories, the literature was reviewed in order to find sources of knowledge for research assessment, and eight examples of such sources were described. These sources of knowledge were conceptualized in a hierarchical structure. For example, knowledge is in the mind of scientists, who are research collaborators, and sometimes play the role of research decision makers. Scientists, or researchers, have their career trajectories registered in curriculum vitae, which represents their scholarly experiences, and their publications (i.e., accomplishments) are among these experiences. At the end of this hierarchical structure are two main types of databases for research assessment: the bibliographical databases and the CV databases, which is exemplified by the Brazilian *Lattes* database ([lattes.cnpq.br](http://lattes.cnpq.br)) in this thesis. The keywords at the right column of Frame 11, express the terminology found in literature that characterizes each one of these eight sources of knowledge.

Frame 11 – Summary of the section 2.4

Sources of Knowledge	Subsection	Keyword
	Scientists / Researchers	Investigators    Experimenters Analyzers    Inquirers Examiners
	Research collaborators	Researchers    work together co-authors scientific collaborators
	Career trajectories	Experiences over time Research career
	Curriculum Vitae	“ <i>a course of study</i> ” “ <i>course of life</i> ” Researcher career trajectory
	Research publications	Books    Book sections    Letters Reviews    Articles Papers    Journal articles
	Bibliographical databases	Scientific literature SCI database
	CV databases	<i>Curriculum Vitae</i> Brazilian <i>Lattes</i> database

Source: The author, 2017.

In conclusion to this investigation, it could be highlighted as sources of knowledge, the concepts of scientists, research collaborators, and research decision makers, as well as, the relevance of the curriculum vitae, and the study of researcher career trajectories.

*Scientists*, for William Whewell (1794-1866), are researchers trained in science at the university, followed a scientific method, and that can receive funding as member of some scientific society.

*Research collaborators* include those who work together in research projects, who are co-authors, or who make substantial contribution for a key step of a scientific research (KATZ; MARTIN, 1997).

*Research decision makers* are those responsible for plan strategies to evaluate and conduct science (e.g., BOZEMAN; GAUGHAN; YOUTIE, 2014; LARGENT; LANE, 2012; THOMSON-REUTERS, 2008).

The term “*Curriculum vitae*”, for academic researchers, is a historical document declared by themselves, which registers their career trajectory (e.g., CAÑIBANO; BOZEMAN, 2009; DIETZ et al., 2000; GAUGHAN; BOZEMAN, 2002).

The study of *researcher career trajectories* has been motivated by the ubiquitous utilization of CVs, and hence strengthened of CV databases (CAÑIBANO; BOZEMAN, 2009; CNPQ, 2017; LANE, 2010; PACHECO et al., 2006). The Brazilian *Lattes* database (lattes.cnpq.br), which will be adopted in this thesis, is acknowledged as one of the best quality information resources in S&T, by providing “*high-quality of data*” (Lane, 2010).

## 2.5 KNOWLEDGE ENGINEERING

Artificial intelligence (AI) is “*an umbrella term for the science of making machines smart*”, as said by the Royal Society (2017, p.122). The term, artificial intelligence, was coined in 1956 during the Dartmouth workshop, and it is related to “*systems that think like humans, act like humans, think rationally, or act rationally*” (RUSSELL; NORVIG, 1995). The fundamentals of AI are in a set of disciplines that contributed with its main ideas, viewpoints and techniques, for instance, philosophy, mathematics, economics, neuroscience, psychology, computer engineering, control theory, cybernetics, and linguistics (RUSSELL; NORVIG, 1995). Nowadays, AI impacts people’s life and society through the creation of emerging technological approaches such as, computer

vision, natural language processing, and knowledge representation and reasoning, among others (STANFORD UNIVERSITY, 2016).

From the view point of technology, AI is supported by Machine Learning, which thought algorithms stablish the interaction between people and computer systems, for instance, when people use voice recognition systems, or when researchers extract knowledge from a vast amount of data exploring experiments in big data (ROYAL SOCIETY, 2017).

In the first generation of knowledge-based systems (KBSs), when AI was in its early days, the development of KBSs was restrict to generic methods to solve problems (e.g., rapid prototyping, symbolic representation, and rules). Those generic methods limited the creation process to small KBSs. The need to change this traditional paradigm originated a new field of study within the scope of AI, which is called Knowledge Engineering (KE) (MARK; SIMPSON, 1991; SCHREIBER et al., 2000; STUDER; BENJAMINS; FENSEL, 1998).

Thus, from the perspective of knowledge representation, AI is supported by Knowledge Engineering (KE), which addresses the construction of the KBSs based on a process of knowledge modelling (SCHREIBER et al., 2000; STUDER; BENJAMINS; FENSEL, 1998; WIELINGA; SCHREIBER; BREUKER, 1992). In this new paradigm, an abstract model is built to better understand the requirements of an area of interest (i.e., a domain knowledge) (SANTOS; TRAVASSOS, 2016).

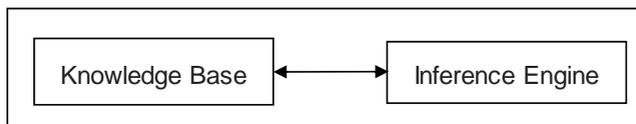
In this section, I review elements of AI that are directly related to KE. I start by introducing the concept of KBS. Then, I present CommonKADS, a methodology to understand the requirements of a KBS. After that, I present Case-based reasoning as a methodology that can be used to implement KBSs. In addition, basic concepts of machine learning are briefly introduced. At the end of the section, a summary underlines the main findings of this review.

### **2.5.1 Knowledge-based systems (KBSs)**

Knowledge-based systems (KBSs), also known as expert systems, are categorized by Hopgood (2005), as a class of artificial intelligence systems that represent the knowledge through declarative techniques (e.g., rule-based, frame-based, model-based, and case-based), in contrast to computational intelligence techniques that represent knowledge through numerical models (e.g., neural networks and evolutionary algorithms). In a KBS there is an explicit separation between knowledge and inference. The knowledge module is called knowledge

base and the control module is called inference engine. A simple view of a KBS is demonstrated by Figure 10.

Figure 10 – The essential components of a KBS

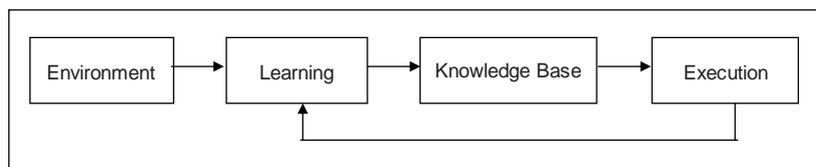


Source: The author, 2016.

The knowledge base contains declarative information (e.g., facts, rules, or relationships) about the problem to be solved, and the inference engine uses the knowledge that is explicitly represented in the knowledge base. This process is in improvement upon previous processes, in which the knowledge was within the structure of the program (HOPGOOD, 2005).

However, these forms of declarative knowledge representation did not suffice because the process of knowledge acquisition depended on experts on specific domains, and also on rules clearly described (HOPGOOD, 2005). Moreover, the choice of an appropriate technique has to take into account aspects, such as, difficulties in modifying the knowledge base and expanding the knowledge representation (XUE; ZHU, 2009).

Figure 11 – Learning system basic structure



Source: Adapted of Xue and Zhu (2009, p.272)

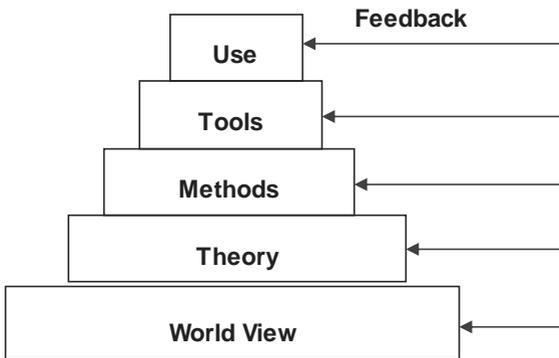
Under those circumstances, one way to overcome such problems could be an efficient learning system (see Figure 11), which automatically learns and adjusts itself according to a set of example solutions available on the knowledge base (HOPGOOD, 2005; XUE; ZHU, 2009). The quest for approaches to solve this problem led to changes in paradigm of KBS construction, by introducing KE methodologies, for example, the CommonKADS.

## 2.5.2 The CommonKADS methodology

CommonKADS (SCHREIBER et al., 2000) emerged in 1995 from the need for a methodology to support a modern KE, by proposing models that went beyond technical requirements of a KBS (PLANT; GAMBLE, 2003).

According to Schreiber et al. (2000), the core structure of the CommonKADS methodology is based on five elements used to construct a KBS – (i) world view, (ii) theory, (iii) methods, (iv) tools, and (v) use, as illustrated by Figure 12. The world view is the first layer of the methodology, which is based on principles that lead to the comprehension of the KBS context (e.g., the goals, mission, values, and priorities). The second layer is theory, which concerns scientific fundamentals necessary to the proposed models. The third layer is methods, these are methodological procedures used to the propose solutions (i.e., models of life cycle, guidelines and techniques for elicitation of knowledge). The fourth layer is tools, which are used to apply the methods (i.e., languages of programming). The fifth layer on the top of the pyramid is use, which represents the experiences with the use of the methodology (i.e., Diagnose and e-Commerce systems). The feedback flows down to each layer along the pyramid.

Figure 12 – The Methodological Pyramid.

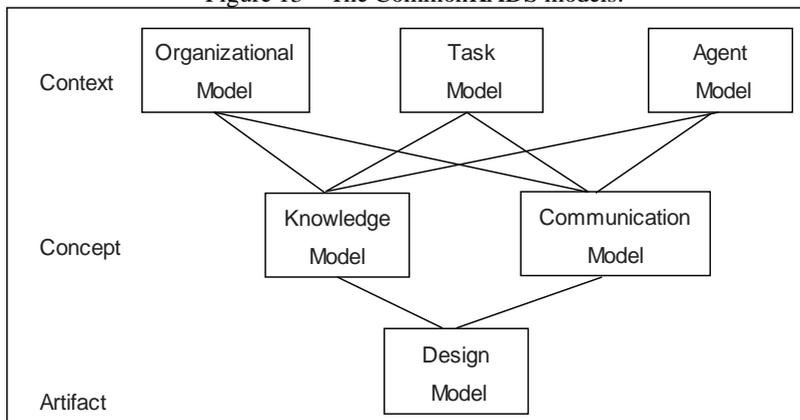


Source: Adapted of Schreiber et al. (2000, p. 15)

The knowledge modelling process of CommonKADS (SCHREIBER et al., 2000) is composed of three layers: Context, Concept and Artifact, as shown in Figure 13. The models of the context layer

investigate the organizational environment in which a KBS will operate. The concept layer addresses the knowledge required for the KBS to solve a particular task. The artifact layer specifies the system architecture and computational mechanisms of the KBS.

Figure 13 – The CommonKADS models.



Source: Adapted of Schreiber et al. (2000, p. 18)

In order to follow these set of models, knowledge engineers must firstly understand the organizational context and the environment in which the KBS will be inserted (i.e., the goals, characteristics of organization, and critical factors for a KBS); Second, they must describe the conceptual knowledge required to solve each task of the KBS (i.e., activities that contribute to achieve the goals); and third, they must describe the technological resources needed to implement the artifacts on the KBS (NAZÁRIO; DANTAS; TODESCO, 2014; PLANT; GAMBLE, 2003; SCHREIBER et al., 2000).

The organizational model, in the context layer, is responsible for identifying problems, opportunities, solutions, and the knowledge that is in the structure of an organization (e.g., processes, tasks, people, and resources). It supports the top-level analysis, including the study of the viability of proposed solutions for the development of the KBS. The other two models will describe the tasks and agents responsible for executing them (SCHREIBER et al., 2000).

The knowledge model is the most important model of the concept layer. It is composed of three essential categories: the domain knowledge, the inference knowledge, and the task knowledge, which detail the structures of knowledge used to perform the tasks identified in the

organizational model. The first, domain knowledge, describes conceptual definitions such as types, rules and facts about an application domain, as in a data model. The second, inference knowledge, contains associations between the concepts presented by the domain knowledge. The third, task knowledge, describes the goals to be achieved, and the strategies that will be adopted by applying the knowledge identified in the previous models (SCHREIBER et al., 2000).

The design model specifies the architecture and computational mechanisms to represent the KBS in a software environment (SCHREIBER et al., 2000).

The CommonKADS methodology distinguishes two main groups of tasks: *analytic tasks* and *synthetic tasks*. Each group is divided in types of problem to be solved by each task, as illustrated in Figure 14.

*Analytic tasks* concern the capability of analysis, especially involving thinking or reasoning. For example, the task of classification analyses a set of objects to establish the correct class for them; Monitoring analyzes a process to find out whether it behaves according to expectations; and Prediction analyses the behavior of a current system to be able to use this knowledge to develop a system in the future (SCHREIBER et al., 2000).

Figure 14 – The groups and types of knowledge intensive tasks.

<b>Knowledge Intensive task</b>	<b>Analytic task</b>	Classification
	Related to use of analysis, logical reasoning, and analogy.	Assessment
		Diagnosis
		Monitoring
		Prediction
		<b>Synthetic task</b>
	The skilled in putting together	Modeling
		Planning
		Scheduling
		Assignment

Source: Adapted of Schreiber et al. (2000, p. 125)

*Synthetic tasks* have the capacity to propose a system that fulfills a set of requirements given. For example, in the task of configuration design, the goal is to assemble components in such way that satisfies predefined requirements.

### **2.5.3 The Case-based reasoning (CBR) methodology**

Case-based reasoning (CBR) is a technique in artificial intelligence that solves problems by using or adapting solutions from old problems (RIESBECK; SCHANK, 2013). It has demonstrated significant efficiency for management and representation of knowledge, and it has been considered not only an artificial intelligence technique, but also a generic methodology for solving problems (AAMODT; PLAZA, 1994; EL-SAPPAGH; ELMOGY, 2015; RICHTER; WEBER, 2013; WATSON, 1999). For example, Kocsis et al. (2014) adopted CBR for mathematical modelling due to its capacity and easy application in tasks of knowledge formulation, knowledge acquisition, and knowledge maintenance.

Taking in consideration CBR as a methodology, I organized this literature review having in mind the five elements required in a methodology, such as those of the methodological pyramid of Schreiber et al. (2000), which consists of world view, theory, methods, tools, and use, as described previously on Section 2.5.2.

#### *2.5.3.1 The CBR world view*

The world view of CBR relies on a basic principle “*to solve a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation*” (AAMODT; PLAZA, 1994, p.2).

#### *2.5.3.2 The CBR theoretical basis*

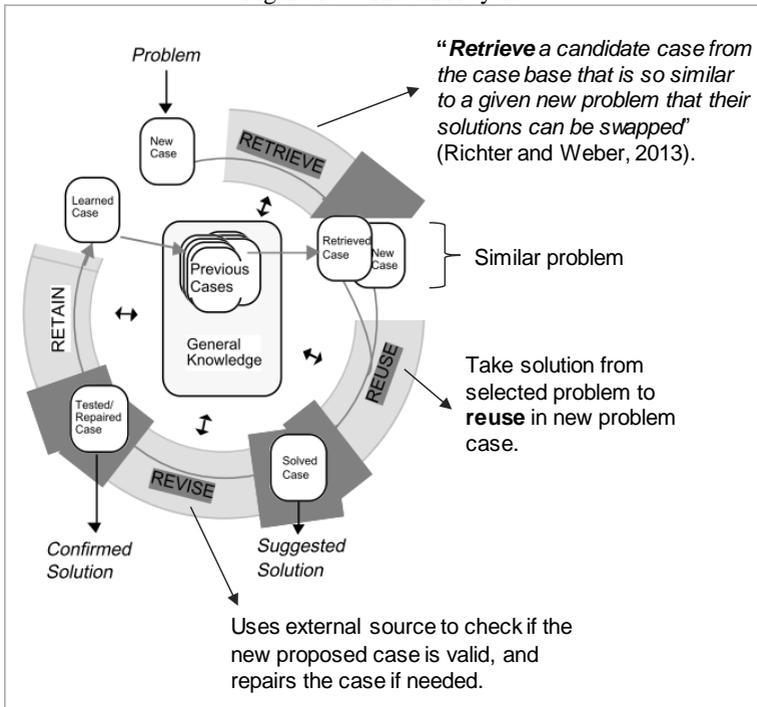
CBR is the result of the intersection of several disciplines based on cognitive science and computer science, which gave CBR its theoretical basis to provide a computational model based on artificial intelligence techniques, in an approach that is similar to human reasoning (RICHTER; WEBER, 2013).

### 2.5.3.3 The CBR method

The CBR cycle (AAMODT; PLAZA, 1994) consists of four basic steps: *Retrieve*, *Reuse*, *Revise*, and *Retain*, which represent the core of CBR methods. Taken into account that a case represents experience, the CBR cycle process incorporates “*what to do*” in order to find useful experiences and how to apply them once they were found (RICHTER; WEBER, 2013). Figure 15 illustrates the CBR cycle.

The initial description of a problem (i.e., a new case) concerns representing the knowledge that is inside the cases through an appropriated data structure. In this task, the problem that originated the needs of a user must be formalized; the cases must be represented; and the CBR knowledge model must to be structured (AAMODT; PLAZA, 1994). The problem can be acquired by performing a dialogue with the user or by using a specific standardized form (RICHTER; WEBER, 2013).

Figure 15 – The CBR cycle



Source: adapted of Aamodt and Plaza (1994, p.8)

This process may sound quite abstract and subjective, however, cases can be represented using AI techniques, which El-sappagh and Elmogy (2015) categorize as traditional (e.g., feature vector, textual, object oriented) or semantical (e.g., ontologies, XML, and OWL) representation methods.

The *Retrieve step* of the CBR cycle is the essence of the CBR methodology. According to Aamodt and Plaza (1994), this step is partitioned in four tasks: The first one consists on understanding a new problem within its context, through the identification of its most relevant features. The second task is to search for a set of plausible candidates, and then in a third task, called initial match, to assess the degree of similarity between the new problem and previous problems stored in a case base. The fourth task is to select the best match from the set of similar cases, and this process is performed by calculating a degree of proximity to the best match.

The *Reuse step* of the CBR cycle tries to reuse in a new problem a solution obtained from the retrieved case solutions. In this task, a solution suggested by the similar case can be combined and reused in the new problem (MÁNTARAS et al., 2005).

The *Revise or Adapt step* in the CBR cycle is necessary to better fit the new problem, whenever significant differences exist between the new problem and the similar case (MÁNTARAS et al., 2005).

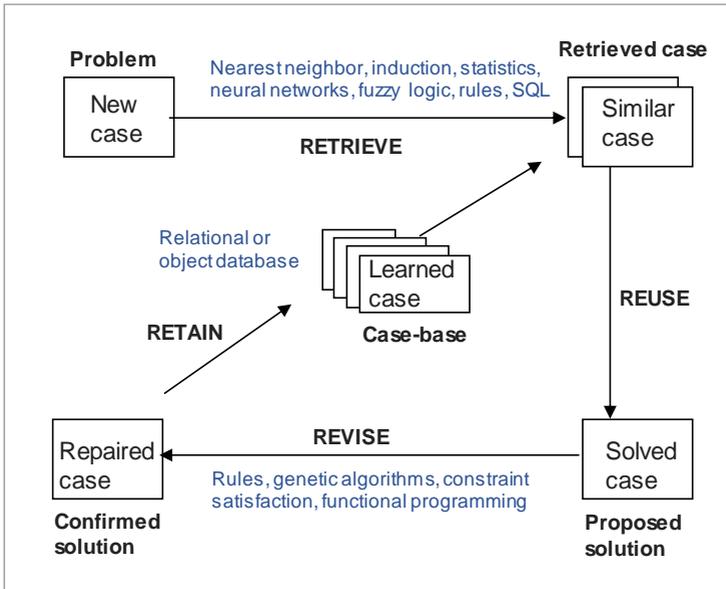
The *Retain step* closes the CBR cycle. After validating the new solution, the useful experience is retained into the system's knowledge for future reuse (AAMODT; PLAZA, 1994).

#### 2.5.3.4 *The CBR techniques*

CBR methodology is able to combine different machine learning techniques and approaches in each step of its cycle (AHMED; BEGUM; FUNK, 2012; WATSON, 1999), as show Figure 16.

In contrast to other artificial intelligence technologies, such as logic programming, neural networks, fuzzy logic and genetic algorithms, CBR does not have its own algorithms to execute the CBR cycle. For example, CBR can use algorithms based on approaches such as nearest neighbor, decision tree, and fuzzy logic to measure the similarity to a target case (WATSON, 1999).

Figure 16 – The CBR cycle combined with Machine learning techniques



Source: Adapted of Watson (1999, p.308)

Figure 16 highlights machine learning technologies used by the CBR methodology, according to the illustration of Watson (1999). Some of such techniques will be detailed on the section 2.5.4 - Machine Learning.

### 2.5.3.5 The CBR definitions and terminologies

In order to synthesize the CBR theory, in this subsection I gather a series of definitions, concepts and terminologies found in the study of Richter and Weber (2013), which I present as following.

**Cases** are experiences, which include problems and solutions in a context, and can be represented by using *feature-value pairs*.

**Definition 1** (RICHTER; WEBER, 2013, p. 34)

*Positive experiences (cases) implement successful solutions*

*Negative experiences (cases) implement failed solutions*

**Definition 2** (RICHTER; WEBER, 2013, p. 38)

For a given set  $U$  of objects, an attribute  $A$  assigns to each object  $O \in U$  some value taken from a set  $dom(A)$ , the domain of  $A$ .

An *attribute-value description* is a finite vector of attributes.

**Definition 3** (RICHTER; WEBER, 2013, p. 39)

*A case base (CB) is a collection of cases.*

**Flat attribute-value representations** are sequential and linear case bases where each case is an attribute-value vector.

**Definition 4** (RICHTER; WEBER, 2013, p. 104)

Suppose an arbitrary set objects  $U$ ,

An attribute  $A$  has domain (also called type)  $dom(A)$  which can be an arbitrary set.

An attribute  $A$  assigns to each  $a \in U$  an element  $A(a) \in dom(A)$ .

Each object from  $U$  is represented as a vector  $(a_1, \dots, a_n)$  with  $a_1 \in dom(A_1)$ ,  $\dots$ ,  $a_n \in dom(A_n)$ .

**Set representations** (RICHTER; WEBER, 2013, p. 118)

Cases in the set have something in common that allows a compact representation by avoiding repetitions. The representation of generalized cases uses the following description elements:

The problem space  $P = P_1 \times P_2 \times \dots \times P_n$  uses  $n$  problem attributes.

The solution space  $S = S_1 \times S_2 \times \dots \times S_m$  uses  $m$  solution attributes.

A query is of the form  $q = (q_1, \dots, q_n)$  using the problem attributes.

A case is of the form  $c = ((p_1, \dots, p_n), (s_1, \dots, s_m))$  using the problem attributes.

### Similarity

The similarity between two cases can be represented with *attribute-value pairs* entails the concept of similarity between attributes and the relative relevance of each attribute.

#### Definition 5 (Richter and Weber, 2013, p. 42)

Be  $CB$  a set of objects and  $p$  be an object; then some  $s$  of  $CB$  is a *nearest neighbor* to  $p$  if there is no object in the  $CB$  that has a higher similarity to  $p$  than  $s$ .

#### Definition 6 (Richter and Weber, 2013, p. 129)

A similarity measure for a problem space  $P$  is a function  $\text{sim}: P \times P \rightarrow [0, 1]$ .

### The nearest neighbour concept

#### Definition 7 (Richter and Weber, 2013, p. 129)

For a given problem  $P$ , a *nearest neighbour* is a problem  $P'$  that has maximal similarity among the problems in  $P$ .

#### Definition 8 (Richter and Weber, 2013, p. 127)

For some fixed  $x$ , each  $y$  that satisfies  $R(x, y, z)$  for all  $z$  is called a *nearest neighbour* of  $x$ .

Notation:  $NN(x, y)$ .

That means that  $NN$  is a relation and the *nearest neighbour* is not necessarily uniquely defined; there may be several such (equally similar) elements. This notion is extended to the *first  $k$ -nearest neighbours*.

The *first  $k$ -nearest neighbours* to  $x$  are listed in a sequence according to their neighbourhood ranking.

Notation:  $NN_k(x) = (z_1, \dots, z_k)$ .

**The Local-Global Principle for Similarity Measures** can be exemplified, according to Richter and Weber (2013, p. 141) by a number of simple measures, and for these authors, each object or concept  $A$  can be described by some construction operator  $C$  from the local elements  $A = C(A_i \ / \ i \in I)$ .

The **global measures** compares objects from a global point of view, and can be the form:

$$Sim(a, b) = \sum_{i=1}^n w_i \cdot sim_i(a_i, b_i), 1 \leq i \leq n$$

where,

$w$  is the weight that reflects the relevance of the attribute on the global measure.

$$\sum_{i=1}^n w_i = 1$$

and  $sim_i(a_i, b_i)$  is the local measure.

#### 2.5.3.6 The CBR use

Similar to the CommonKADS methodology, the CBR methodology can be applied in analytical tasks or synthetical tasks. Despite this categorization, CBR applications have been developed by using elements of both categories for different areas, such as, medical diagnosis, technical diagnosis, e-commerce, time-series analysis, and recommender systems, among others, such as listed in Frame 12.

Frame 12 – The use of CBR methodology on related studies and applications

Related Study (PhD Thesis)	Application Area	Type of Task	Description
(CECI, 2015)	Sentiment analysis	Analytical	The study combines CBR methodology and domain ontology as a strategy to identify the sentiment aggregated to a particular statement.

(Cont.)

<b>Related Study (PhD Thesis)</b>	<b>Application Area</b>	<b>Type of Task</b>	<b>Description</b>
(GUNDERSEN, 2014)	Streaming data	Analytical	The study applies CBR on streaming data to solve problems of real-time decision making.
(GUNAWARDENA, 2013)	Recommender systems	Analytical	The study identifies collaboration opportunities that will have better chances of success.
(AGORGIANITIS et al., 2016)	Planning of business process	Synthetical	The work investigates CBR applications specialized on business workflow monitoring.
(KOCISIS et al., 2014)	Design and Scheduling	Synthetical	The study proposes a decision support system based on CBR applied on scheduling problems.

Source: The author, 2017.

Case-based reasoning (CBR) is also associated to cognitive processes including analogical reasoning, which learn from reminding past problems that can be adapted to help solve a new problem, and this is the fundamental principle of CBR (MÁNTARAS et al., 2005). In learning-by-analogy process, the knowledge about a fact is retrieved from memory; after that, it can be appropriately transformed, applied to the new situation, and then, stored for future use (CARBONELL; MICHALSKI; MITCHELL, 1983).

#### **2.5.4 Machine Learning methods**

Machine Learning (ML) is a field of AI that allows computer systems to learn directly from examples, data, and experience. It establishes the interaction between people and computer systems through methods created by the intersection of computer science, statistics, and data science (ROYAL SOCIETY, 2017).

More precisely, the goal of ML is to explore learning methods applicable in different types of domain knowledge. Such methods are based on cognitive science and simulate human learning. The most common examples of learning methods are: Learning from instruction; Learning by analogy; Learning from examples, Explanation-based learning, Learning by deduction, and Inductive learning (CARBONELL; MICHALSKI; MITCHELL, 1983; XUE; ZHU, 2009).

Furthermore, machine learning deals with the identification of common characteristics in data, by applying several techniques, such as classification and clustering (e.g., HALL, 1999; ROYAL SOCIETY, 2017). In Classification, or supervised learning, a system is trained with data that were previously labelled, that is, data are classified into one or more classes providing an immediate feedback. The system learns how these data are structured, and attempts to predict the category of a new object. In contrast to classification, Clustering or unsupervised learning is an approach that uses data which has not been labelled, then, the system is trained to decide which objects should be grouped together.

In the next subsections, I present some of the most important Machine Learning algorithms, categorized on supervised and unsupervised learning. At the end of this section, I present some basic concepts of descriptive statistics.

#### 2.5.4.1 *Supervised Learning Algorithms: Feedback Algorithms*

**Decision trees** are the most popular algorithms used in classification tasks, which are based on a tree data structure which is composed of nodes (i.e., features), branches (i.e., associated values of features) and leaves (i.e., classes) (HALL, 1999). The classical studies of Quinlan (1992, 1986) on induction of decision trees describe the ID3 algorithm (i.e., Iterative Dichotomiser) and its successor, the C4.5 algorithm. The difference between them is that C4.5 estimates the error rate of every subtree, and if there is a lower error rate, this subtree can be replaced or discarded. The C4.5 algorithm is implemented as the J48 algorithm by the Weka tools (<http://www.cs.waikato.ac.nz/ml/weka/>).

**Linear regression algorithms** are based on statistic regression analysis, and they are widely applied on numeric predictions. The simple idea is to express a class of a set of data as a linear combination of attributes, with predetermined weights (WITTEN; FRANK, 2005).

**Probabilistic algorithms** make the assumption that feature values are statistically independent given a class, and estimates the probability of attribute values (i.e., instances) within the class to decide in which class belongs this instance. One example of such approach is the Naïve Bayesian classifier, which is adopted in supervised inductions to representing, using, and learning probabilistic knowledge (JOHN; LANGLEY, 1995).

**Lazy learning algorithms** according to Aha (1998) are lazy because of the three characteristics: Defer, Demand-Driven, and Discard. Firstly, all training data are stored and deferred until each new example is compared to those that are stored (WETTSCHERECK; AHA; MOHRI, 1997); Second, the information is replied by combining information stored and the training data (AHA, 1998); Third, the constructed query and intermediate results are discarded (AHA, 1998).

The *k-nearest-neighbor classifier (k-NN)* is an example of the lazy learning algorithms (WETTSCHERECK; AHA; MOHRI, 1997). The studies on k-NN originated in the early 1950s, it was adopted as a classification method in the 1960s, and achieved popularity through the studies of Aha in the 1990s, who proposed a framework and methodology called instance-based learning algorithms (IBL) (e.g., AHA, 1998, 1992; AHA; KIBLER; ALBERT, 1991). In this thesis, the concept of nearest neighbour (NN) was stated on section 2.5.3, based on the definition of Richter and Weber (2013, p. 127-129), which is: “*For a given problem P, a nearest neighbour is a problem P' that has maximal similarity among the problems in P*”.

The *Instance-based learning (IBL)*, as described by Aha et al. (1991), extend the k-NN algorithm, and are methods, like the case-based reasoning (CBR), they assume that “similar instances have similar classifications” (AHA; KIBLER; ALBERT, 1991, p. 41). The IBL algorithms use a distance metric to compare new instances with existing ones, and the closest is then used to assign the class of the new instance (WITTEN; FRANK, 2005). The first IBL algorithm proposed by Aha et al. (1991) was the IB1 algorithm, which is “identical to the k-NN algorithm except that it normalizes its attributes' ranges, processes instances incrementally, and has a simple policy for tolerating missing values” (AHA; KIBLER; ALBERT, 1991, p. 42). In the next versions, evolutions in the IB1 algorithm based the IB2 and IB3 algorithms. According to Aha et al. (1991), the main advantage to applying the IBL approaches is their simplicity.

The *Relief algorithm* is another example inspired on instance-based learning methods, which was proposed by Kira and Rendell (1992). According to the authors, it is a feature weighting based approach, and addresses the classification tasks of identifying features statistically relevant to the target problem. A distance function  $diff$  is computed between two instances to find the nearest neighbors. An extension of the Relief, called ReliefF algorithm, was proposed by Knonenko, Robnik-Sikonja, and Pompe (1996) to solve some limitations of the first version, for instance, the problem of dealing with more than two classes. The result of Relief and ReliefF algorithms are ranks of attributes by weigh that reflect their relevance (HALL, 1999).

#### 2.5.4.2 *Unsupervised Learning Algorithms (Clustering algorithms)*

***K-Means*** is classified as a partition method, in which the elements of a set are grouped in clusters. The process starts with the choice of  $k$  initial points to represent initial cluster centers (i.e. the means). Then, the data points are grouped around each  $k$ -mean, which is being recalculated along the process (WITTEN; FRANK, 2005). One disadvantage of this method is that it requires specifying the number of clusters (MAKKAR, 2015). To minimize this problem some extensions of  $k$ -Means have been proposed to estimate the number of clusters, for instance, the X-Means method (PELLEG et al., 2000).

***Density-based Spatial Clustering of Application with Noise (DBSCAN)*** (ESTER et al., 1996) is a class of clustering called density-based method. In the DBSCAN, clusters are modeled as dense regions separated by sparse regions. The method does not require the input of a specific number of clusters. However, two input parameters are necessary: the minimum number of points (i.e., MinPts) required to form a cluster, and the Eps-neighborhood ( $\epsilon$ ) a threshold value of distance that delimits the neighbourhood area.

***Expectation maximization (EM)*** (DEMPSTER; LAIRD; RUBIN, 1977) is a probability-based clustering method that attempts to find the distribution probability of an object belonging to each cluster. The method is divided in two steps, expectation that calculates the cluster probabilities, and maximization that estimates the likelihood of the distributions given a set of data (WITTEN; FRANK, 2005). The cluster membership filter is an extension of the EM algorithm implemented in

the Weka tools (<http://www.cs.waikato.ac.nz/ml/weka/>), which is a filter that uses a density-based clusterer to generate cluster membership values, i.e., the probability of each instance being classified in a class or another (WITTEN; FRANK, 2005).

### 2.5.4.3 *Typical difficulties found in Machine Learning classification tasks*

In the Machine Learning (ML) systems, the goal is to understand the “*real-world*” to simulate it through computational models (JORDAN; MITCHELL, 2015; ROYAL SOCIETY, 2017; XUE; ZHU, 2009). Thus, machine learning systems try to realize the context of a problem by learning from its characteristics, through for example, any approach provided by a supervised learning method (HALL, 1999; WITTEN; FRANK, 2005). However, at least three difficulties have been found in this task of learning: The first is the lack of negative instances; the other is in the identification of attribute relevance of a problem; and another consists in the selection of the best threshold of a solution, which is common in retrieval applications. In the next sub-sections I will describe these three concerns, which demand particular solutions.

#### *The lack of negative instances*

In classification tasks, the goal is to train a classifier which accurately predicts its category, given a new unlabeled example (HALL, 1999; WITTEN; FRANK, 2005). For instance, a binary classifier must predict if a new example is fit or unfit for a purpose. A common problem occurs in retrieval applications, when a subset of data contains only positive examples, and the rest of the dataset remains unlabeled (GUNAWARDENA; WEBER; STOYANOVICH, 2013; IENCO; PENSA, 2016; LIU et al., 2003). The lack of negative instances makes traditional methods of classification learning inapplicable, because a dataset with examples labeled as positive and negative is essential for a classifier (e.g., a binary classifier) (LIU et al., 2002). In order to solve this problem some approaches have been proposed, such as the studies of Liu et al. (2003, 2002) and Gunawardena et al. (2013) which will be described in the following paragraphs.

Liu et al. (2003, 2002) address the problem of building text classifiers with only positive and unlabeled documents, and these authors call this problem *partially supervised classification*. In this case, there are not negative documents, and the positive are mixed with unlabeled

documents. Thus, it is not possible to know which documents are positive or negative. The study proposes a method, called the *Spy technique in S-EM* (see Figure 17), which using the *EM algorithm* (DEMPSTER; LAIRD; RUBIN, 1977), defines the probability a document is positive or negative.

Figure 17 – The Spy technique in S-EM

1.  $N = NULL;$
2.  $S = \text{Sample}(P, s\%);$
3.  $U_s = U \cup S;$
4.  $P_s = P - S;$
5. Assign each document  $d_j$  in  $P$  the class  $c_1$ ;
6. Assign each document  $d_j$  in  $U$  the class  $c_2$ ;
7. Run EM(MS,P);
8. Classify each document  $d_j$  in MS;
9. Determine the probability *threshold*  $t$  using S;
10. For each document  $d_j$  in  $M$
11.     if its probability  $Pr[c_1][d_j] < t$
12.          $N = N \cup \{d_j\}$
13.     else  $U = U \cup \{d_j\};$

Source: Adapted of Liu et al. (2003, 2002)

The method simulates the idea of sending some “spy” documents from the set  $P$  (i.e., documents labeled as positive) to observe the set  $M$  (i.e., documents labeled as mixed). The goal is to investigate the behavior of unknown positive documents that are mixed with unlabeled documents, in an attempt to identify those that have some characteristic that makes them different from the others.

The proposed approach randomly selects 10% of the documents from the *positive set*  $P$  as spies, which are denoted by  $S$ , and includes them inside the *set*  $M$  (i.e.,  $P = P - S$  and  $M = M \cup S$ ). In the next step, the *EM algorithm* (DEMPSTER; LAIRD; RUBIN, 1977) is applied over the *set*  $M$  to calculate the most likely probability that a document is negative (class 0) or positive (class 1). A *threshold* ( $t$ ) is employed to take this decision, in which, documents in  $M$  with lower probabilities than  $t$  are the most likely negative documents, denoted by  $N$ . On the other hand, documents in  $M$  that have higher probabilities than  $t$  become unlabeled documents, denoted by  $U$ . At the end, 15% of documents with probability lower than  $t$  are classified as *negative* ( $N$ ). The process is iterative, and in

the beginning of each iteration, the spy documents  $S$  from the last one, are put back to the *positive set P*.

Gunawardena et al. (2013) also poses the problem of the lack of negative instances, and proposes an approach that uses clustering algorithms. The approach is applied in a CBR problem, and it seeks to identify cases that are well aligned versus cases that are poorly aligned. The intuition of these authors is that the difference between well and poorly aligned cases is revealed when outliers (i.e., objects that are not close to any cluster) are identified in the problem and solution spaces. For them, cases where similar problems have similar solutions are well aligned, and on the contrary this premise, the cases are poorly aligned. Thus, for such authors, the challenge is to determine which cases are poorly aligned. This approach is also inspired in the Spy technique proposed by Liu et al. (2003, 2002), and it will be used posteriorly to learn weights, in that poorly aligned cases are understood as potential negative instances.

### *The relative relevance of attributes of a problem*

Other quite expressive problem in classification tasks relies on the matter of the quality of attributes, in other words, in the discernment that some attributes are more relevant than others (ROBNIK-ŠIKONJA; KONONENKO, 2003; WITTEN; FRANK, 2005; YU; LIU, 2004). This problem concerns to studies on the feature selection and the feature weighting approaches.

In *feature selection approaches*, a set of relevant features are identified and those irrelevant or redundant are removed from the data set. Among the advantages of this process is the reduction of the dimensionality of the data, which allows learning algorithms to operate faster. It also improves the accuracy on classification tasks, and hence, imply in a subset of attributes that are sufficient to describe a target concept (HALL, 1999; KIRA; RENDELL, 1992). However, the contra point is that feature selection assigns binary weights to features, which suggest whether the attribute should be retained or discarded (ROBNIK-ŠIKONJA; KONONENKO, 2003; WETTSCHERECK; AHA; MOHRI, 1997).

*Feature weighting* can be considered an extension of feature selection, in that, it emphasizes the importance of each attribute, by assignment of weights in an individual evaluation process, in which is given to them different degrees of relevance (WETTSCHERECK; AHA; MOHRI, 1997; WITTEN; FRANK, 2005; YU; LIU, 2004). In this

approach, none of attributes is discarded (WETTSCHERECK; AHA; MOHRI, 1997).

Several approaches have been proposed to solve both, the *feature selection* and the *feature weighting* problems. Such approaches produces a rank of attributes, which can be categorized by two different criteria: individual attribute evaluation or subset attribute evaluation (HALL; HOLMES, 2003). Example of such approaches are: *Correlation Based Feature Selection (CFS)* (HALL, 1999), *Information Gain (IG)* (QUINLAN, 1992), *Gain Ratio (GR)* (QUINLAN, 1992), *Symmetrical Uncertainty* (PRESS et al., 1988), *Relief* (KIRA; RENDELL, 1992), *ReliefF* (KNONENKO; ROBNIK-SIKONJA; POMPE, 1996).

### ***Selecting the best threshold of a solution***

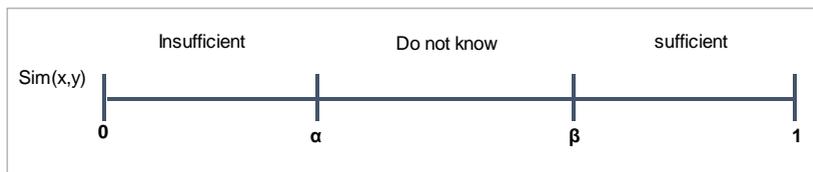
A problem in machine learning applications is the choosing of a solution that best represents a set of candidate solutions (WEBER-LEE et al., 1996). For instance, this is a common problem in approaches based on *k-NN algorithms*, whose goal is to identify the maximal similarity among a set of objects (i.e., exemplars), but this process became difficult whether many equally similar objects are returned (RICHTER; WEBER, 2013; WETTSCHERECK; AHA; MOHRI, 1997). Another problem, occurs in the application of the *Spy technique* (LIU et al., 2002), over a set of data that consists only positive documents. Then, a threshold is calculated to decide which documents are most likely to be negative. In the following, three solutions these two problems are briefly described.

Weber-Lee et al., (1996) poses the problem of searching for an appropriate threshold as a difficult task in the CBR retrieval tasks (i.e., to select the best match), and propose an approach based on geometrical fuzzy clustering algorithms, called Fuzzy c-means. In this approach three measures are calculated, based on the data from the solutions offered by the retrieved cases, to compute the threshold. These measures are inspired on the *Theory of Typicality* (FRIEDMAN; MING; KANDEL, 1995), and they are: *The Most Typical Value (MTV)*, the *Most Typical Deviation (MTD)*, and the *Definite Typical Value (DTV)*.

Richter and Weber (2013) also suggests computing degrees of similarity and dissimilarity, to obtain intervals of thresholds (e.g., insufficient, do not know, and sufficient), as shown in Figure 18, the defuzzification process. For example, if the similarity is in the “*insufficient*” area, then the solution is not accepted; If it is in the “*sufficient*” area, then it is accepted. But, if it is in the “*do not know area*”

additional criteria for a decision is needed, because this is an uncertainty area.

Figure 18 – The defuzzification process.



Source: Richter and Weber (2013, p. 131)

Liu et al., (2002) proposes a simple solution to determine the *threshold* ( $t$ ), to take the decision, in which, documents with lower probabilities than  $t$  are the most likely negative documents. The approach first sorts the documents according to their probability. After that, a selected noise level of  $t\%$  is used to decide  $t$  (e.g.,  $t = 5\%$ ,  $10\%$ ,  $15\%$ ,  $20\%$ ). In Liu et al., (2002) is used  $t = 15\%$ .

#### 2.5.4.4 Basic statistical measures

In this subsection, I present some of the most common statistical measures that are used in combination with machine learning classification methods, and will be used along of this study. A list of some such measures is shown 3.

Frame 13 – Common statistical measures used in machine learning

Measure	Description
Min	The minimum value of all values in an interval of values.
Max	The maximum value of all values in in an interval of values
Measures of central tendency	Describes the tendency around a particular value
Mean	The arithmetic Mean of all values in an interval of values. It is calculated as the sum of all values divided by the number of values of the interval, and indicates the center of a distribution.

(Cont.)

Measure	Description
Median	The middle value within an interval of values. It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not.
Mode	The most frequently occurring value in an interval of values
Measures of dispersion	Describes the extent in which the values are spread out from the average
Range	The difference between the highest and lowest values in an interval of values.
Standard deviation	The deviate from the mean in an interval of values. It is commonly used when the mean is used as measure of central tendency.
Variance	The square of the standard deviation
Quartiles	The data is divided into four equal parts. Quartiles is used when the median is used as the measure of central tendency.

Source: adapted from Nicholas (1999)

In addition to these measures, special attention will be given to two essential concepts provided by the statistics field, which are broadly related to machine learning methods. One is the concept of normalization, and the other is the concept of accuracy.

*Normalization* is a method that may be used before a classification process, and it is required to equalize ranges of the features from different scales, in order to obtain the same proportion between them, making features comparable (SINGH; VERMA; THOKE, 2015). For example, if the feature is many times larger than another, both features may be brought to the same proportion, and they can be considered equivalent.

Several techniques have been proposed to implement normalization (e.g., Min-Max normalization, and Z-Score normalization), and many studies have investigated the relation between choosing the appropriated normalization technique to improve the performance of classification accuracy (e.g., JAYALAKSHMI; SANTHAKUMARAN,

2011; SINGH; VERMA; THOKE, 2015). 4 lists two examples of normalization techniques.

Frame 14 – Examples of normalization techniques

<b>Normalization technique</b>	<b>Description</b>
Min-Max normalization	<p>Min-Max Normalization rescales the feature values from one range of values to a new range of values, preserving exactly the same proportion between the values.</p> $X_{i,0,1} = (x_i - x_{min}) * \frac{(x_i - x_{min})}{(x_{max} - x_{min})} + x_{min}$ <p>Where,</p> <ul style="list-style-type: none"> <li><math>x_i</math> is each data point i</li> <li><math>x_{min}</math> is the minimum value among all features</li> <li><math>x_{max}</math> is the maximum value among all features</li> <li><math>X_{i,0,1}</math> is the feature normalized between 0 and 1</li> </ul>
Z-Score normalization	<p>The Z-Score normalization uses the statistical mean and standard deviation, to compute a normalized measure that entail a more centralized set of data, with zero being the central point. The follow equation defines the Z-Score normalization.</p> $X_{i,0,1} = \frac{x_i - (x_{max} + x_{min})/2}{(x_{max} - x_{min})/2}$ <p>Where,</p> <ul style="list-style-type: none"> <li><math>x_i</math> is each data point i</li> <li><math>x_{min}</math> is the minimum value among all features</li> <li><math>x_{max}</math> is the maximum value among all features</li> <li><math>X_{i,0,1}</math> is the feature normalized between 0 and 1</li> </ul>

Source: Adapted of Etzkorn (2011), Jayalakshmi and Santhakumaran (2011), and Singh et al. (2015)

**Accuracy**, according to the dictionary, is “*the ability to work or perform without making mistakes*” (e.g., <http://www.merriam-webster.com>). Similarly in machine learning, the correctness of a classification can be evaluated by computing its accuracy (e.g., SOKOLOVA; LAPALME, 2009). Thus, classification accuracy is defined as the percentage of examples correctly classified by an algorithm, which requires the application of several measures to achieve this end (e.g., HALL, 1999; SOKOLOVA; LAPALME, 2009).

Frame 15 presents the *confusion matrix*, which is composed of four indicators as described next.

- True positive (tp) is the number of correctly recognized class examples.
- *False positive (fp)* is the number of examples that were incorrectly assigned to the class
- *False negative (fn)* is the number of examples that were not recognized as class examples.
- *True negative (tn)* is the number of correctly recognized examples that do not belong to the class

Frame 15 – The confusion matrix for binary classification accuracy

Data classification	Classified as positive	Classified as negative
Positive	True positive (tp)	False negative (fn)
Negative	False positive (fp)	True negative (tn)

Source: Adapted of Sokolova and Lapalme (2009)

These four indicators are applied on measures to compute accuracy in binary classifications, as shown in Frame 16.

Frame 16 – Measures to compute accuracy in binary classifications

Measure	Description	Equation
Accuracy (A)	Accuracy (A) is the overall effectiveness of a classifier	$A = \frac{tp + tn}{tp + fn + fp + tn}$

(Cont.)

Measure	Description	Equation
Precision	Precision represents the number of correctly classified positive examples divided by the number of examples labeled by the system as positive.	$Precision = \frac{tp}{tp + fp}$
Recall	Recall is the number of correctly classified positive examples divided by the number of positive examples in the data.	$Recall = \frac{tp}{tp + fn}$

Source: Adapted from Sokolova and Lapalme (2009) and Hall (1999)

Another statistical method to estimate accuracy of learning algorithms is the *cross-validation*, which divides the dataset into two subsets, one used to train a model and the other to validate the model (REFAEILZADEH; TANG; LIU, 2009). The most basic form of cross-validation is the *k-fold-cross-validation*, in that a *dataset (D)* is randomly split into *k mutually exclusive subsets* (i.e., the folds) of approximately equal size, and to compute the overall accuracy, the overall number of correct classifications is divided by the number of instances in the dataset (HALL, 1999; KOHAVI, 1995; REFAEILZADEH; TANG; LIU, 2009). In this method, after the dataset have been randomly divided into *k* equal subsets, *k* rounds of learning are performed, and on each round, *1/k* of the data is held out as a test set and the remaining examples are used as training data (RUSSELL; NORVIG, 1995).

A special case of the *k-fold-cross-validation* widely adopted is the *Leave-One-Out Cross-Validation (LOOCV)* that uses a single split of the data into the folds, by making *k = n*, where *n* is the size of the dataset. Thus, LOOCV removes one instance from the data at each iteration and submits it as a new unclassified instance using the remaining instances to classify the one left out (GU; AAMODT, 2006; KOHAVI, 1995; REFAEILZADEH; TANG; LIU, 2009).

### 2.5.5 Concluding remarks

In summary, I gathered the main findings of the literature review on Knowledge Engineering (KE), based on Frame 17, which synthetizes the subjects investigated in each subsection, in keywords. Section 2.5 addressed three main aspects of KE, its goals, methodologies, and

methods used to implement Knowledge-based systems (KBSs). These aspects will be briefly described below.

Frame 17 – Summary of the section 2.5

Knowledge Engineering	Subsection	Keyword
	Knowledge-based system (KBS)	Artificial Intelligence Expert systems Machine Learning systems
	CommonKADS methodology	Context    Organizational model Concept    Knowledge model Artifact    Design model
	CBR methodology	Artificial Intelligence CBR world view    CBR techniques CBR cycle CBR use    CBR terminologies
	Machine Learning methods	Artificial Intelligence Computer science Supervised learning Unsupervised learning Classification    Clustering    Statistics

Source: The author, 2017.

*Knowledge Engineering* is a field of Artificial intelligence (AI) originated from the need to construct KBSs. It is based on a process that, before implementing, requires the understanding, modelling and representation of knowledge of a particular field of study (MARK; SIMPSON, 1991; SCHREIBER et al., 2000; STUDER; BENJAMINS; FENSEL, 1998). The quest for approaches to better design a KBS entailed the development of methodologies, such as, the CommonKADS (SCHREIBER et al., 2000).

The knowledge modelling process of CommonKADS is composed of three layers: *Context*, *Concept* and *Artifact*. The context layer investigates the organizational environment in which a KBS will operate; the concept layer addresses the knowledge required for the KBS to solve a particular task; and the artifact layer specifies the system architecture and computational mechanisms used to implement the KBS (SCHREIBER et al., 2000).

After investigating CommonKADS as a methodology to design a KBS, I investigated other methodologies to implement KBSs, and I found in Case-based reasoning (CBR), characteristics that go beyond those commonly found in AI methods. Thus, I reviewed the literature considering the five elements required in a methodology, such as those of the methodological pyramid of Schreiber et al. (2000), which are: world view, theory, methods, tools, and use.

The fundamental principle of CBR is based on analogical reasoning, which learns from remembering past problems that can be adapted to help solve a new problem (MÁNTARAS et al., 2005). The CBR cycle (AAMODT; PLAZA, 1994) is the core of CBR, and is a process composed of four steps: Retrieve, Reuse, Revise, and Retain. In this process, through analogical reasoning, the knowledge about a fact is retrieved from memory; after that, it can be appropriately transformed, applied to the new situation, and then, stored for future use (CARBONELL; MICHALSKI; MITCHELL, 1983).

In contrast to other artificial intelligence technologies, such as logic programming, neural networks, fuzzy logic and genetic algorithms, CBR does not have its own algorithms, but, a set of Machine Learning (ML) methods that can be combined to implement the four steps of the CBR cycle (WATSON, 1999). Thus, this entailed the study of ML methods that could be adopted in the proposed method.

Machine Learning (ML) is a field of AI that establishes the interaction between people and computer systems through methods created by the intersection of computer science, statistics, and data science (ROYAL SOCIETY, 2017). Such methods are categorized in two types: Classification (supervised learning), and clustering (unsupervised learning). In the first, a system is trained with previously labelled data, and then, a new instance of data can be classified into one or more classes, predicting its category. In contrast, unsupervised learning is an approach that uses data which has not been labelled, then, the system is trained to decide which objects should be grouped together.

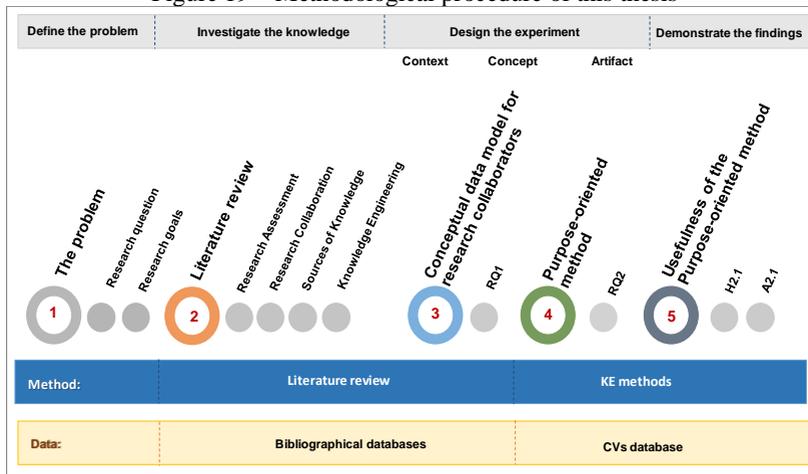
In addition to ML methods, I also investigated approaches to solve three common problems in CBR applications: the lack of negative instances when applying supervised learning; the identification of relative relevance of attributes; and the selection of the best threshold of a set of solutions. Also, statistical measures that are used in combination with ML methods, such as normalization and accuracy, were investigated.

### 3 METHODOLOGY

#### 3.1 INTRODUCTION OF THE CHAPTER

This section describes the set of methodological procedures used in this thesis. In the scientific inquiry, the set of methodological procedures is called scientific method, which is composed of a cycle that consists of the following elements: Observations, questions, hypotheses, experiments, and generalizations (e.g., BHATTACHERJEE, 2012; GIGCH, 1979). This thesis applies a set of methodological procedures based on the scientific method and knowledge engineering (KE), as illustrated in Figure 19.

Figure 19 – Methodological procedure of this thesis



Source: The author, 2017.

Following the scientific method, this thesis firstly defines the problem and formulates the research questions and goals; after that the knowledge that bases the experiments is investigated in the literature; the next step is to design the experiment to answer the research questions; at the end, the findings are demonstrated. This process is developed in five steps. Along the process, specific methods are applied, and data are used to demonstrate the experiment. The next subsection describes the methodological procedures.

### 3.2 METHODOLOGICAL PROCEDURES

This subsection outlines the five steps that compose the methodological procedures of this thesis: Step 1: The problem; Step 2: The literature review; Step 3: The conceptual data model for research collaborators; Step 4: The purpose-oriented method; Step 5: The usefulness of the purpose-oriented method.

Defining the theme was the first task of the study, which is assessing researcher quality for collaborative purposes. This task was executed before I defined the problem. It was firstly investigated through a previous literature review on “*research collaboration*”, which was presented and approved on the “*2014 Painel Científico*”, an internal conference of the PPGEGC/UFSC, in March 2014. Thus, in the step1 of the methodological procedure the problem, its contextualization, relevance and originality are presented.

When presenting the problem, I also considered what scientific approach would be the most appropriate to address the investigation. For being a theme related to procedures for measuring, assessing and quantifying evaluations, this kind of research follow a *quantitative world view*, in which, “*variables can be measured, typically on instruments, so that numbered data can be analyzed using statistical procedures*” (CRESWELL, 2009, p.4).

This research is also characterized as an *applied research*, which is a type of investigation “*directed primarily towards a specific, practical aim or objective*” (OECD, 2015b, p.365), in contrast to a pure basic research, which does not seek “*long-term economic or social benefits or making any effort to apply the results to practical problems*” (OECD, 2015b, p.378).

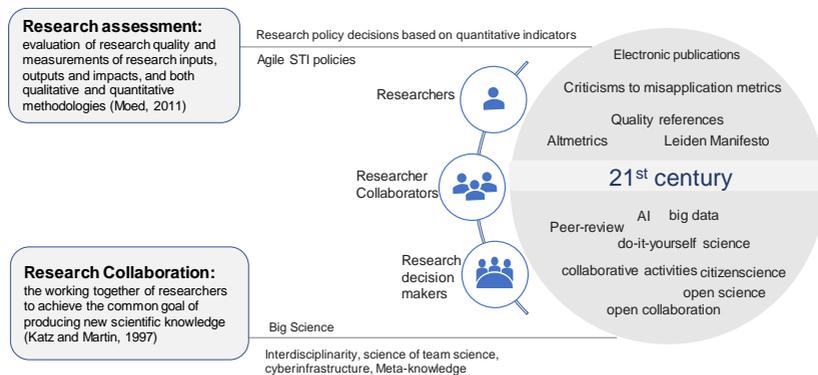
Furthermore, this thesis focuses on the *Interdisciplinary collaboration*, which was previously stated in Section 2.3.3, and according to Mobjörk (2010, p.868), “*Interdisciplinarity is characterized by the collaboration between researchers from different disciplines, with a common methodological approach and a shared problem* “. This thesis encompasses aspects of the *Multidisciplinary collaboration*, however, it does not address aspects of *Transdisciplinarity*, such as the collaboration of researchers outside academia.

In step 2, this thesis reviews the literature, by using data from bibliographical databases, in order to scrutinize four constructs: Research Assessment (see Section 2.2), Research Collaboration (see Section 2.2), Sources of Knowledge for Research Assessment (see Section 2.3), and Knowledge Engineering (see Section 2.4).

The steps 3 and 4 constitute the experimental design, in which a *purpose-oriented method* is proposed to solve the problem identified in step 1. In order to design the experiment, the CommonKADS methodology and the CBR methodology are applied. The first was used for modelling the knowledge of the proposed method, and the second was used for implementing the proposed method. Apart from these methodologies, machine learning methods (see Section 2.5.4) were applied to build the *purpose-oriented method*.

The CommonKADS is composed of three layers: *Context*, *Concept* and *Artifact*. The context layer has its main core in the organizational model, which represents the contextual environment in which a KBS operates (see Section 2.5.2). I illustrate in Figure 20, the environment within which the proposed purpose-oriented method is inserted, inspired by the CommonKADS organizational model, and based on the findings produced by the literature review.

Figure 20 – The organizational environment in which the proposed method will operate



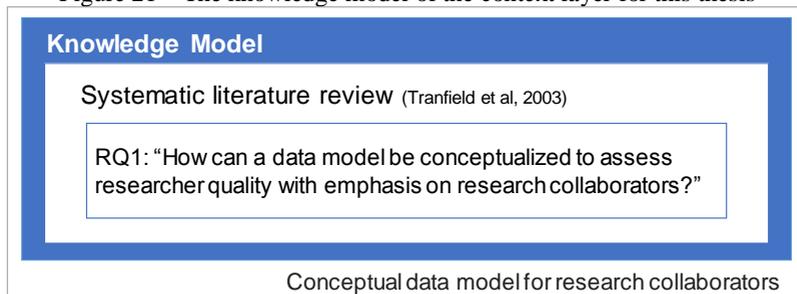
Source: The author, 2017.

Assessing research quality for collaborative purposes is the task to be performed, and the picture illustrates the 21<sup>st</sup> century organizational environment in which this task is inserted. This environment is influenced by elements such as technology, bibliometric indicators, and a new way to conduct research aligned with the Big Science concept. There are three agents interacting with this environment, researchers, research

collaborators, and research decision makers, who are sources of knowledge for research assessment.

Step 3 describes the *concept layer* whose knowledge model is the most important. It details the domain knowledge, that is, structures of knowledge such as types, rules and facts about an application domain, for instance, those described in a data model (SCHREIBER et al., 2000). In this step, researchers and research collaborators are investigated in deep, as unit of analysis of the assessment. For this, a systematic literature review, based on Tranfield et al (2003) is conducted to gather structures of knowledge that characterizes researchers and research collaborators, to answer the first research question: “*How to conceptualize a data model to assess researcher quality with emphasis on research collaborators*”, as illustrated in Figure 21.

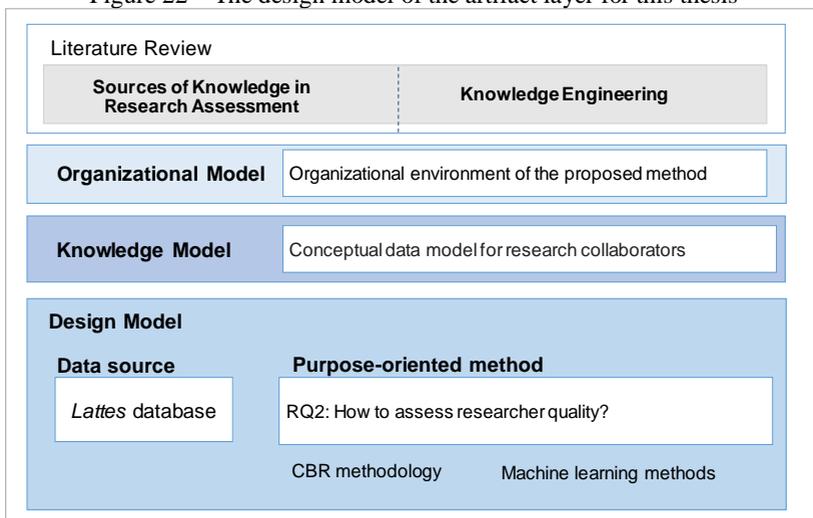
Figure 21 – The knowledge model of the context layer for this thesis



Source: The author, 2017.

The *artifact layer* is represented by the design model, which specifies the architecture and the computational mechanisms required to represent the KBS in a software environment (SCHREIBER et al., 2000). Thus, in step 4, the second research question is answered, “*RQ 2: How to assess researcher quality?*”. In order to achieve this goal, a *purpose-oriented method* is designed, supported by the CBR methodology and machine learning methods. In this thesis, the CommonKADS design model refers to step 4, which apart from the proposed *purpose-oriented method*, includes the specification of CV database utilized in the experiments. Figure 22 illustrates how the CommonKADS design model is applied in this thesis.

Figure 22 – The design model of the artifact layer for this thesis



Source: The author, 2017.

Finally, in the step 5, the usefulness of the proposed method is demonstrated through two experiments, as described in the next paragraphs.

The first experiment demonstrates hypothesis “*H2.1: A purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method*”. This experiment is aligned with the second principle of the Leiden Manifesto (HICKS et al, 2015), which states that performance should be measured against the research missions of the institution, group or researcher. The approach adopted in this experiment integrates purpose with information about the selection process that needs to assess researcher quality. It uses a sample of 100 CVs data from the Brazilian Lattes database (lattes.cnpq.br) and the method adopted to validate the results is Leave-One-Out Cross-Validation (LOOCV).

The second experiment is based on the assumption A2.2, in which, “incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment”. This experiment is also aligned with the Leiden Manifesto (HICKS et al, 2015), and concerns at least four of its principles. The second principle, as mentioned above, suggests aligning the metrics with the mission and purposes of the institutions; the third principle emphasizes the need to acknowledge local instead of universal research;

the sixth principle recommends taking into account the specificities of fields and publication practices; and the seventh principle suggests that individual researchers should be assessed based on a qualitative judgement of their portfolio.

In order to execute the second experiment, an application scenario is created by using data from the Brazilian Lattes database (CNPq, 2017), including researchers from the Microcephaly Epidemic Research Group (MERG) (<http://www.cpqam.fiocruz.br/merg/>). This experiment is divided into two parts: In the first part, the proposed method does not consider the career trajectory of researchers. In the second part, the treatment of career trajectories is incorporated into the purpose-oriented method, and at the end, the results of both parts are compared.

It is important to point out that more details about the methods, techniques, and data used by each experiment will be given in Chapter 7.

### 3.3 CONCLUDING REMARKS

In Chapter 3 the methodological procedures used in this thesis are presented. The thesis relies in a quantitative world view, and it is characterized as an applied research. Furthermore, it focuses on a interdisciplinary collaboration approach.

This thesis is conducted in five steps: *Step 1*: The problem; *Step 2*: The literature review; *Step 3*: The conceptual data model for research collaborators; *Step 4*: The purpose-oriented method; and *Step 5*: The usefulness of the purpose-oriented method.

In step 1, the problem is defined, and the research questions are formulated, and in step 2, the literature is reviewed in order to scrutinize four constructs: *Research assessment*, *research collaboration*, *sources of knowledge for research assessment*, and *Knowledge Engineering*.

The next two steps address the experimental design, by applying two KE methodologies to propose the *purpose-oriented method*. The CommonKADS methodology is used to modelling the knowledge of the proposed method, and the CBR methodology is used for implementing the proposed method. Step 3 describes the *concept layer* of CommonKADS methodology. For this, a systematic literature review, based on Tranfield et al (2003) is conducted to investigate the domain knowledge about researchers and research collaborators. The result of this step is the *conceptual data model for research collaborators*.

Step 4 presents the *artifact layer* of the CommonKADS methodology, which specifies the architecture and the computational mechanisms required to design the proposed *purpose-oriented method*.

Thus, the CBR methodology and Machine Learning methods are used to this task, which includes the specification of the CV database utilized in the experiments.

Finally, in the step 5, the usefulness of the proposed method is demonstrated through two experiments. The first experiment introduces the *purpose-oriented method*, by using a sample of 100 CVs data from the Brazilian *Lattes* database ([lattes.cnpq.br](http://lattes.cnpq.br)) and the method adopted to validate the results is *Leave-One-Out Cross-Validation* (LOOCV).

The second experiment effectively demonstrates the *purpose-oriented method*, through an experimental scenario that includes researchers from the *Microcephaly Epidemic Research Group* (MERG) (<http://www.cpqam.fiocruz.br/merg/>). This experiment is divided into two parts: In the first part, the proposed method does not consider the career trajectory of researchers. In the second part, the treatment of career trajectories is incorporated into the purpose-oriented method, and at the end, the results of both parts are compared.

The next sections present the steps 3 to 5 of this methodological procedure. Chapter 4 concerns the *conceptual data model for research collaborators*. Chapter 5 describes the data used in the experiments; Chapter 6 proposes the *purpose-oriented method*; and Chapter 7 demonstrates the usefulness of the proposed method.



## 4 CONCEPTUAL DATA MODEL FOR RESEARCH COLLABORATORS

### 4.1 INTRODUCTION OF THE CHAPTER

In this chapter, through the methodology proposed by Tranfield et al. (2003), a systematic literature review is conducted, in order to search for attributes used to characterize researchers with a particular emphasis on collaborative work. The reason to investigate attributes of researchers is that my ultimate goal is to conduct studies to inform research related decisions, primarily focusing on researcher quality assessment.

The review investigates data from the research collaborator's organizational context, as suggested by Bozeman, Fay and Slade (2013) and Lane et al. (2015). Furthermore, the study is concerned with the use of CV data, whose investigation is crucial to understanding what components in a researcher's profile are valued in the literature.

This systematic review focuses on descriptors of individual research collaborators, and not on those that describe relationships between different research collaborators, because the goal is to support studies that can analyze one individual researcher collaborator at a time. Apart from that, this study prioritizes attributes commonly used by studies on research quality assessment, which can be objectively valued and that are likely available in profiling systems.

The results of this study suggest a conceptual data model that is backed by the literature in the field. This conceptual model represents the knowledge model of the CommonKADS methodology, which details the domain knowledge, that is, structures of knowledge such as types, rules and facts, about an application domain (SCHREIBER et al., 2000).

In the next sections, the systematic review adopted in the study is outlined, and its results presented with special focus on the conceptual data model; at the end, a concluding remarks section ends chapter 4.

### 4.2 SYSTEMATIC LITERATURE REVIEW

The systematic literature review applied in this study was introduced by Tranfield et al. (2003), it has been widely adopted (e.g., PERKMANN et al., 2013) and recommended for doctoral level studies (ARMITAGE; KEEBLE-ALLEN, 2008). In this study, it is implemented throughout three stages: *Stage I Planning the review*, *Stage II Conducting the review*, and *Stage III Reporting the review*. Next, each one of these stages will be described.

### 4.2.1 Stage I: Planning the review

The objective of first stage is to establish the focus of the review, and this is done through the first research question (i.e., RQ1: “*How to conceptualize a data model to assess researcher quality with emphasis on research collaborators?*”). The research question guides the investigation by defining the scope. Then, to limit this scope a research strategy is defined through the parameters, such as, the studies that were included, search period, and initial keywords. The review focus on the search for attributes to characterize scholarly researchers, with a particular emphasis on collaborative work in order to conceptualize the data model.

The search strategy is defined taking into account the relevance of the articles to be included in the review. Firstly, it is limited to literature in research collaboration. The review include data from the research collaborator’s organizational context, as suggested by Bozeman, Fay and Slade (2013) and Lane et al. (2015). Secondly, it emphasizes the importance of the groundwork of the three pioneer authors, Derek de Solla Price, Donald Beaver, and Eugene Garfield. It also includes articles which have at least one reference to Katz and Martin (1997), given their seminal contribution in defining collaboration. In addition, the recent technologies that changed the environment of scientific practices in the beginning of the new century are also considered to limit the scope to studies published from the year 2000 to 2014.

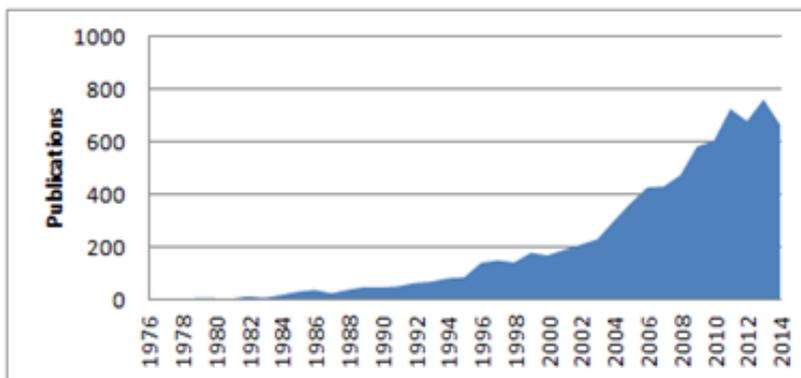
The parameters of search are then defined in five steps. Step 1 establishes the parameters of search protocol. They include sources of information (e.g., articles, books, digital libraries), search period (i.e., start date), language of publication, and initial keywords. Step 2 defines criteria for selecting articles based on relevance. Step 3 identifies articles based on relevance by reading titles and abstracts. Step 4 selects the resulting set of articles used as source of analysis, by reading the full texts of the articles of them. Step 5, the attributes that characterize research collaborators of each resulting article are analyze.

### 4.2.2 Stage II: Conducting the review

Stage II is when the steps described above are executed. In Step 1, the SCOPUS was chosen as the main digital library (see Section 2.4.6). Other parameters of the search protocol were established to limit to journal articles, reviews, and conference papers; and to the English language. The search was conducted to find publications that included at least one of the following expressions: “*scientific collaboration*” or “*science*

*collaboration*" or *research collaboration*" or *collaborative science*", or *collaborative research*" or *research collaborator*" or *researcher collaborator*" or *collaborative researcher*" or *scientific collaborator*" or *science collaborator*". Step 1 was executed in April 2014 and resulted in 6988 publications, from 1976 to 2014, as presented in Figure 23.

Figure 23 – Growth of scientific collaboration from 1976-2014.



Source: The author, 2016.

Step 2 is when the relevance criteria to filter results are used, keeping only the ones that reference at least one of the pioneer authors of research collaboration. The search is also limited to publications by recency because of our interest in collaborations that became more frequent with technological advances such as regular use of email and internet, post year 2000. This filtering reduced the set to 916 publications.

In the Step 3, the relevance of the articles was assessed by reading their titles and abstracts, and 101 publications was selected. In Step 4, the 101 publications are read and in 37 of them, was found relevant attributes for this study, these constitute the resulting review set. In Step 5, each of the 37 articles are analyzed, and the attributes of research collaborators are collected.

#### 4.2.3 Stage III: Reporting the review

The reporting is organized in two parts. Firstly, a quantitative analysis of the results is presented based on the implementation of steps 1 to 4. This analysis includes the main authors, in which journals they published, and the most used keywords in these articles. Secondly, a

qualitative analysis of the attributes found in the articles of the resulting review set is describe, which is organized in a data model to assess researcher quality. This qualitative analysis will be presented in the section 4.3. This data model answers the research question RQ1: “*How to conceptualize a data model to assess researcher quality with emphasis on research collaborators?*”.

#### 4.2.3.1 Authors

Table 1 lists the nine authors identified with the highest number of publications in the resulting review set. The second column presents the number of publications for each author. The third column shows the number of citations each author received within the articles in the resulting review set. These authors are originally from six different countries, namely, Italy, Iran, Malaysia, Spain, UK, and USA.

Table 1 – Most frequent authors included in this study.

Author	Number of Publications	Citations within the resulting review set	City/Country
Abramo, Giovanni	5	4	Italy
Beaver, Donald deB.	2	15	USA
Boardman, Craig	2	8	USA
Bozeman, Barry	5	13	USA
Corley, Elizabeth	4	10	USA
Didegah, Fereshteh	3	6	UK
Gazni, Ali	3	6	Iran
Jonkers, Koen	2	1	Spain
Thelwall, Mike	3	2	UK

Source: The author, 2016.

Considering these authors, this review highlights in special the studies of two of them, Donald Beaver and Barry Bozeman. The first, Donald Beaver, was one of the pioneering authors in research collaboration as aforementioned. Professor Beaver investigated the root of growth of co-authorship in a series of three studies called, “*Studies in Scientific Collaboration - Part I, II and III*” (BEAVER; ROSEN, 1978, 1979a, 1979b). Two of his recent studies were selected by this review (BEAVER, 2001, 2004), demonstrating the strong and current influence

of this author in the recent literature on research collaboration. The second author, Barry Bozeman, were the first one to consider the attributes of research collaborators, in Bozeman, Fay and Slade (2013), a literature review.

#### 4.2.3.2 *Journals*

Table 2 lists the 16 different journals included in the resulting review set. It also indicates the number of articles from each journal, among them, this review highlights two of them, *Scientometrics* and *Research Policy*. These journals have the largest number of publications included in this analysis.

Table 2 – Journals included in this study

Journal	Articles
<i>Scientometrics</i>	8
<i>Research Policy</i>	7
<i>Journal of Informetrics</i>	4
<i>Higher Education</i>	2
<i>Journal of the American Society for Information Science and Technology</i>	2
<i>Research Evaluation</i>	2
<i>Administrative Science Quarterly</i>	1
<i>American Sociologist</i>	1
<i>Journal of Socio-Economics</i>	1
<i>Journal of Technology Transfer</i>	1
<i>Project Management Journal</i>	1
<i>Public Administration</i>	1
<i>Science and Public Policy</i>	1
<i>Social Networks</i>	1
<i>Social Studies of Science</i>	1
<i>Studies in Higher Education</i>	1

Source: The author, 2016.

### 4.2.3.3 Keywords

Table 3 lists in order of frequency, the as the most frequent keywords compiled from all publications that are included in the resulting review set.

Table 3 – Most frequent keywords in publications included in this study

<b>Keywords</b>	<b>Frequency</b>
Research collaborations	13
Co-authorship	5
Collaboration patterns	5
Bibliometrics	4
Collaboration	4
Universities	4
Productivity	4
Research	3
Big science	2
Citation impact	2
scientific collaboration	2

Source: The author, 2016.

## 4.3 THE CONCEPTUAL DATA MODEL

In this section, a data model to assess researcher quality that emphasis research collaborators, is proposed. This data model has the purpose of organizing and representing the research collaborators context through multiple dimensions that categorize the various attributes found in the literature review, as illustrated in Figure 24.

Figure 24 – The research collaborator conceptual data model includes the context where a research collaborator delivers scientific activities to achieve accomplishments



Source: Duarte, Weber and Pacheco (2016a)

The model, as previously mentioned, is motivated by the recommendation from Bozeman, Fay and Slade (2013) and Lane et al. (2015) who suggest that studies should include the context in which the research collaborator operates. The dimensions are the researcher, the institutions, the accomplishments, and the career. As illustrated in Figure 24, the first three are dimensions, while the career is a meta-dimension that moves across the others to describe the research collaborator's trajectory.

The proposed model highlights the relations between the dimensions. For example, the research collaborator as a researcher affiliated to institutions, whose accomplishments are achieved along their career. After to categorize the attributes under each dimension, it makes clear that some attributes do not originated from the dimensions, but from the relation between the research collaborator and these dimensions. Abramo, D'Angelo and Murgia (2014), for example, utilize type of affiliation as an attribute that only exists in the relation between a research collaborator and its institution. This attribute does not originate from the researcher nor the institution, but the relation between them.

#### **4.3.1 Scope and limitations of the conceptual data model**

This literature review investigates attributes that describe research collaborators individually. This means that it does not include attributes that relate two collaborators, such as their proximity. Furthermore, the study is restricted to those attributes that are commonly available in data sources, such as from research profiling systems and other publicly available data and whose values can be obtained in a reasonably objective fashion. In that case, this work does not include studies that focus exclusively on subjective aspects such as those studied in cognitive science and psychology.

In order to preserve fairness to the original authors of the articles whose publications are included in the resulting review set, and in reason of the interpretation of some attributes might be influenced by the categories under which they were classified, the classification of these attributes under the categorical dimensions presented in the conceptual data model is of exclusive responsibility of the author of this thesis and are not to be attributed to the original authors.

Along these lines, I noticed that the relevance of some attributes selected from the review is implicit in the articles. For example, Bozeman, Fay and Slade (2013) explicitly propose *gender* as a personal attribute of a research collaborator. On the other hand, Abramo, D'Angelo, and Di

Costa (2014) do not explicitly propose *academic rank* as an attribute to characterize a research collaborator. They use these attributes while studying collaboration patterns across academic ranks, and when using it, they assign values to research collaborators (ABRAMO; D'ANGELO; DI COSTA, 2014).

Besides these limitations, there might be some considerations to ethical use of the attributes described here. Some limitations may originate in cultural aspects. Particularly considering data from CVs, some cultures require them, while others prohibit the inclusion of attributes such as marital status or age, and therefore their use should consider contextual and local ethical rules.

### 4.3.2 The attributes of research collaborators

This section describes the attributes found in the literature review. These attributes are then categorized within the researcher, institutions, achievements, and career dimensions. Approximately sixty attributes are categorized in these four dimensions of the research collaborator's context.

#### 4.3.2.1 *Researcher*

The dimension researchers stems from the fact that a research collaborator is a researcher. For this reason, the research collaborator inherits personal attributes typically used to describe researchers. The dimension "*researchers*" includes attributes that are easily available in researchers' CVs. Examples are personal attributes such as *name* and *age* (e.g., ABRAMO; D'ANGELO; DI COSTA, 2009; BOZEMAN; FAY; SLADE, 2013). Other set of attributes included in this dimension are field of training, tacit knowledge and network ties, which Bozeman, Fay and Slade (2013) called human capital attributes.

In addition, this study would like to highlight two articles namely Jeong et al. (2014) and Maglaughlin and Sonnenwald (2005) which propose subjective attributes that are neither available nor can be objectively collected (e.g., motivational factors). The latter article suggests subjective attributes in the context of interdisciplinary collaborations and proposes attributes such as learning and teaching, new discoveries, fun, and external rewards as values. However, as a limitation of the model, attributes that are considerate subjective are not include in the list. Frame 18 provides the entire list of attributes under this dimension

with their pertinent references that were collected from the resulting review set.

Frame 18 – List of attributes characterizing research collaborators as researchers

Attribute	Source
Name	(ABRAMO; D'ANGELO; MURGIA, 2014; CORLEY; SABHARWAL, 2010; GAZNI; THELWALL, 2014)
Age	(BOZEMAN; FAY; SLADE, 2013; DAHLANDER; MCFARLAND, 2013; LEE; BOZEMAN, 2005; SCELLATO; FRANZONI; STEPHAN, 2015)
Gender	(ABRAMO; D'ANGELO; MURGIA, 2013a, 2014; BOZEMAN; FAY; SLADE, 2013; DAHLANDER; MCFARLAND, 2013; HUNTER; LEAHEY, 2008; JONKERS; CRUZ-CASTRO, 2013; LEE; BOZEMAN, 2005; SCELLATO; FRANZONI; STEPHAN, 2015)
Contact information, e.g., email address	(GAZNI; THELWALL, 2014; YOUTIE; BOZEMAN, 2014)
Languages, Marital status, Citizenship, Nationality	(LEE; BOZEMAN, 2005)
Ethnicity	(DAHLANDER; MCFARLAND, 2013)
Country of origin	(SCELLATO; FRANZONI; STEPHAN, 2015)
Country of residence	(CORLEY; SABHARWAL, 2010)
Complementary skills	(BOZEMAN; CORLEY, 2004; MAGLAUGHLIN; SONNENWALD, 2005)
Race, Tacit knowledge, Fields of training, Network ties	(BOZEMAN; FAY; SLADE, 2013)
Tenure status	(BOARDMAN; CORLEY, 2008; DAHLANDER; MCFARLAND, 2013)

Source: Duarte, Weber and Pacheco (2016a)

#### 4.3.2.2 *Institutions and affiliations*

As previously mentioned, research collaborators are members of collaborations (KATZ; MARTIN, 1997) that occur in physical or virtual interactive spaces. These interactive spaces are present in the institutional structure of universities, research centers, and industrial labs, which encourage their researchers to participate in collaborations, both within and across institutions (HUNTER; LEAHEY, 2008). The institutions that researchers are affiliated are thus an important element in the context of

the research collaborator. For this reason, various articles suggest attributes of institutions to be used to characterize research collaborators (e.g., BOARDMAN; CORLEY, 2008; BOZEMAN; FAY; SLADE, 2013; CUMMINGS; KIESLER, 2007). Note that one researcher may be affiliated to multiple institutions through his or her career and some affiliations may be simultaneous.

Analogous to the researcher dimension, some attributes of the institution dimension are available in a researcher's CV, such as the *institution's name* (e.g., KUMAR; JAN, 2013). Others may be easily available in online sources. Examples are *size* and *age of an institution* (e.g., ABRAMO; D'ANGELO; DI COSTA, 2009; BOZEMAN; FAY; SLADE, 2013; CARILLO; PAPAGNI; SAPIO, 2013; KNOBEL; PATRICIA SIMÕES; DE BRITO CRUZ, 2013). Attributes like *size* and *age* contribute to the characterization of *prestige* (e.g., CARILLO; PAPAGNI; SAPIO, 2013), which has been explicitly used as an attribute by Hunter and Leahey (2008). Abramo, D'Angelo, and Di Costa (2009) explain how the *size of institution* may indicate the amount of opportunities available for researchers to collaborate internally. Bozeman, Fay and Slade (2013) suggest that the limited opportunities in smaller institutions may be the cause of increased collaborations with industry.

Frame 19 – List of attributes characterizing research collaborators from the perspective of her or his institutions

Attribute	Source
Name of institutions	(BOARDMAN; CORLEY, 2008; HUNTER; LEAHEY, 2008; KNOBEL; PATRICIA SIMÕES; DE BRITO CRUZ, 2013; KUMAR; JAN, 2013)
Acronyms of institutions	(KUMAR; JAN, 2013)
Size of institution	(ABRAMO; D'ANGELO; DI COSTA, 2009; BOZEMAN; FAY; SLADE, 2013)
Age of institution	(CARILLO; PAPAGNI; SAPIO, 2013; KNOBEL; PATRICIA SIMÕES; DE BRITO CRUZ, 2013)
Institutional prestige, Geographic location	(HUNTER; LEAHEY, 2008)
Infrastructure institutional and funds available, Institutional labor policies	(ABRAMO; D'ANGELO; DI COSTA, 2009)
Number of PhD and post-docs	(CARILLO; PAPAGNI; SAPIO, 2013)

Source: Duarte, Weber and Pacheco (2016a)

Attributes such as *funds* and *infrastructure* (e.g., ABRAMO; D'ANGELO; DI COSTA, 2009) have been utilized despite being potentially more difficult to obtain. *Number of PhD* and *postdocs* is not only hard to obtain, but it is also variable. Carillo, Papagni and Sapio, (2013) comment that *the number of PhD students* and *postdocs* may either increase or decrease productivity depending on whether they provide valuable research assistance or increase their load of responsibilities. They conclude in their study that their presence is positively associated with high quality publications (e.g., CARILLO; PAPAGNI; SAPIO, 2013). Frame 19 lists attributes originated from institutions that were learned from the literature review.

When introducing the conceptual data model for the research collaborator, it was mentioned that the resulting attributes were organized using dimensions and the relationship a researcher may have with the entity of the dimension. The relationship between a researcher and an institution is an affiliation.

Several authors recommended attributes to characterize researchers based on the relation with his or her affiliation. The initial attribute that qualifies the affiliation is the *type of affiliation* (e.g., SCELLATO; FRANZONI; STEPHAN, 2015). Frame 20 lists attributes characterizing research collaborators from the perspective of a researcher's affiliations.

Frame 20 – List of attributes characterizing research collaborators from the perspective of a researcher's affiliations

Attribute	Source
Type of affiliation, Job position	(SCELLATO; FRANZONI; STEPHAN, 2015)
Institutional address	(ABRAMO; D'ANGELO; MURGIA, 2013a, 2013b, 2014; DE STEFANO et al., 2013; PEREZ-CERVANTES; MENA-CHALCO; CESAR JR., 2012)
Academic rank	(ABRAMO; D'ANGELO; MURGIA, 2013a, 2014; DE STEFANO et al., 2013)
Disciplines, fields and subfields	(ABRAMO; D'ANGELO; MURGIA, 2013a, 2014)

Source: Duarte, Weber and Pacheco (2016a)

Some affiliation attributes may seem personal attributes at first, but they originate from the affiliation relation. Examples are *institutional address* (e.g., ABRAMO; D'ANGELO; MURGIA, 2013a, 2013b, 2014;

DE STEFANO et al., 2013; HUNTER; LEAHEY, 2008; KUMAR; JAN, 2013; PEREZ-CERVANTES; MENA-CHALCO; CESAR JR., 2012) and *academic rank* (e.g., ABRAMO; D'ANGELO; MURGIA, 2013b, 2014; DE STEFANO et al., 2013). Although the usual labels for academic ranking in American universities are frequently mentioned (e.g., assistant, associate professor), these values differ depending on the career structure of the institution.

#### 4.3.2.3 *Accomplishments*

The third dimension to characterize research collaborators stems from the accomplishments they achieve. Research collaborators typically work alone or in collaboration with others to achieve accomplishments. This literature review reveals that the characteristics of achieved accomplishments, as well as those of the collaborative processes that research collaborators engage, are relevant to characterize the research collaborator. Frame 21 lists attributes characterizing research collaborators from the perspective of researchers' accomplishments.

Frame 21 – List of attributes characterizing research collaborators from the perspective of researchers' accomplishments

<b>Attribute</b>	<b>Source</b>
Digital Object Identifier (DOI)	(PEREZ-CERVANTES; MENA-CHALCO; CESAR JR., 2012)
Publication title	(DE STEFANO et al., 2013)
Journal title, Publication keywords	(CORLEY; SABHARWAL, 2010; DE STEFANO et al., 2013)
Number of authors	(GAZNI; DIDEGAH, 2011; NEWMAN, 2004)
Number of citations	(ABRAMO; D'ANGELO; MURGIA, 2014; CORLEY; SABHARWAL, 2010; GAZNI; DIDEGAH, 2011)
Journal impact factor	(DIDEGAH; THELWALL, 2013)
Number of foreign countries	(GAZNI; DIDEGAH, 2011)
Epistemic authority	(BEAVER, 2004; BOZEMAN; FAY; SLADE, 2013)
Research grants	(LEE; BOZEMAN, 2005)

Source: Duarte, Weber and Pacheco (2016a)

Despite the various potential types of accomplishments (e.g., journal articles, books, books chapters, grants and patents), the publications in this review do not suggest or imply an attribute to specify

whether an accomplishment is a publication, or a funded proposal. Most attributes directly describe characteristics of accomplishments. For example, *number of authors* (e.g., GAZNI; DIDEGAH, 2011; NEWMAN, 2004) can be used to describe both publications and funded projects. Other attributes are specific to one type of accomplishment, such as *journal impact factor* (e.g., DIDEGAH; THELWALL, 2013) and *journal title* (e.g., CORLEY; SABHARWAL, 2010; DE STEFANO et al., 2013).

Beaver (2004) and Bozeman, Fay and Slade (2013) consider *epistemic authority* as an attribute to describe accomplishments that can be used to characterize its collaborators. Epistemic authority is related to how influential an accomplishment is and has been correlated to the number of citations (BEAVER, 2004).

#### 4.3.2.4 Processes

The result of a collaborative process is the accomplishment that a researcher achieves. Assigning values for attributes to characterize a collaborative process may be difficult because a researcher typically records accomplishments (e.g., an article accepted for publication) and not the process that leads to it (e.g., number of meetings to discuss article draft with collaborators).

Nevertheless, it is in the process to achieve accomplishments that research collaborators engage in collaborations. As attributes in Frame 22 reveal, many authors consider these attributes important when studying collaborations and collaborators (e.g., BOZEMAN; FAY; SLADE, 2013; GAZNI; DIDEGAH, 2011).

A common attribute used to describe the process is the *type of collaboration*, which has been pointed out since the work of Katz and Martin (1997). Authors use *type* to describe whether the collaborative purpose is *institutional*, *domestic*, or *international* (e.g., ABRAMO; D'ANGELO; MURGIA, 2014; DIDEGAH; THELWALL, 2013; GAZNI; DIDEGAH, 2011; IBÁÑEZ; BIELZA; LARRAÑAGA, 2013).

Gazni, Sugimoto and Didegah (2012) utilize *size of the collaboration* as an attribute of the collaborative process, which differs from the attribute from the accomplishment called *number of authors*. As an attribute of the process, the *number of participants* (e.g., CORLEY; BOARDMAN; BOZEMAN, 2006) may include team members that may not have collaborated for the entire duration of the process or staff members who do not have their names added to scientific

accomplishments due to their administrative role. Size of collaboration is more general than number of participants because it can be used to compute other aspects such as the *number of institutions* and other elements to compute *size* (e.g., GAZNI; SUGIMOTO; DIDEGAH, 2012).

Frame 22 – List of attributes characterizing research collaborators from the perspective of the collaborative process delivered to achieve accomplishments

Attribute	Source
Types of collaboration	(ABRAMO; D'ANGELO; MURGIA, 2014; DIDEGAH; THELWALL, 2013; GAZNI; DIDEGAH, 2011; IBÁÑEZ; BIELZA; LARRAÑAGA, 2013; THELWALL; SUD, 2014)
Administrative roles in the collaborative group	(BEAVER, 2001; BOZEMAN; FAY; SLADE, 2013)
Management style within research team	(BOZEMAN; FAY; SLADE, 2013; CHOMPALOV; GENUTH; SHRUM, 2002)
Collaboration Strategies	(LEE; BOZEMAN, 2005)
Size of collaboration	(CARILLO; PAPAGNI; SAPIO, 2013; CUMMINGS; KIESLER, 2007; GAZNI; SUGIMOTO; DIDEGAH, 2012)
Number of institutions	(GAZNI; DIDEGAH, 2011)
Budget	(CORLEY; BOARDMAN; BOZEMAN, 2006)
Duration of collaboration	(BROCKE; LIPPE, 2013; CUMMINGS; KIESLER, 2007; JEONG; CHOI; KIM, 2014)
Participants of collaboration	(CORLEY; BOARDMAN; BOZEMAN, 2006)
Participation incentives	(CORLEY; BOARDMAN; BOZEMAN, 2006)
Communications, physical interaction, physical meetings, informal communications, Communication technologies	(ABRAMO; D'ANGELO; MURGIA, 2013a; BOZEMAN; FAY; SLADE, 2013; JEONG; CHOI; KIM, 2014)

Source: Duarte, Weber and Pacheco (2016a)

Several authors are concerned with the ways of communication in a process, particularly whether communication technologies are used (e.g., ABRAMO; D'ANGELO; MURGIA, 2013a). Aspects such as whether process include *face-to-face meetings* are considered because of studies that investigate their relations with productivity (e.g.,

CUMMINGS; KIESLER, 2005). Frame 22 lists attributes characterizing research collaborators from the perspective of the collaborative process they follow to achieve collaborative accomplishments.

#### 4.3.2.5 Career

The final element we consider is the career of a research collaborator that represents a longitudinal account of an individual's productivity (DIETZ et al., 2000). This concept entails the remaining attributes found in this literature review, as shown in Frame 23.

Given the nature of a career trajectory that describes the historical progress of a research collaborator's path, the values for its attributes are mostly derived from the other dimensions. This justifies the formulation of career as a meta-dimension in the conceptual data model of research collaborators.

Frame 23 – List of attributes that characterize research collaborators from the perspective of their career trajectories.

Attribute	Source
Career age	(BOZEMAN; FAY; SLADE, 2013; LEE; BOZEMAN, 2005; PEREZ-CERVANTES; MENA-CHALCO; CESAR JR., 2012)
Career stages	(BOZEMAN; FAY; SLADE, 2013; DAHLANDER; MCFARLAND, 2013)
Previous collaboration / Experiences	(BOZEMAN; FAY; SLADE, 2013; DAHLANDER; MCFARLAND, 2013; JONKERS; TIJSSEN, 2008)
Effects of seniority	(JONKERS; TIJSSEN, 2008)
International mobility data	(SCCELLATO; FRANZONI; STEPHAN, 2015)
Trajectories	(BOZEMAN; FAY; SLADE, 2013)
Years since PhD	(JONKERS; TIJSSEN, 2008; LEE; BOZEMAN, 2005)
Career productivity, Publications since PhD	(LEE; BOZEMAN, 2005)

Source: Duarte, Weber and Pacheco (2016a)

*Career age* is mentioned by many authors (e.g., BOZEMAN; FAY; SLADE, 2013; LEE; BOZEMAN, 2005; PEREZ-CERVANTES; MENA-CHALCO; CESAR JR., 2012), and it can be simply inferred from a CV. *Career stages* (e.g., DAHLANDER; MCFARLAND, 2013) however vary depending on the culture of the research collaborator, and

on the perspective of who is charged with the task of inferring career stages. Previous collaborations are suggested by Dahlander and McFarland (2013).

Effects of *seniority* (e.g., JONKERS; TIJSEN, 2008) may be considered as an interpretation of one's career stage. This attribute also varies with the culture. On the other hand, *career productivity* can be independent of stages, such as the *number of publications* since a researcher received her or his *doctoral degree* (e.g., LEE; BOZEMAN, 2005).

Scellato, Franzoni and Stephan (2015) study international networks and thus suggest attributes that are specific of international collaborations. One of them is the international mobility, which includes multiple sub-attributes that attempt to describe the trajectory of a research collaborator who has been affiliated with institutions in multiple countries.

#### 4.4 CONCLUDING REMARKS

This chapter focused on attributes that characterize research collaborators. The literature was reviewed and approximately sixty attributes were identified and categorized in four dimensions, *researcher*, *institutions*, *accomplishments*, and *career trajectories*. This context was used to categorize, organize, and describe research collaborator in a conceptual data model, illustrated in Figure 24.

This review favors objective attributes and does not include studies that focus exclusively on subjective aspects such as those studied in cognitive science and psychology. These attributes are of objective nature in that they may be directly available in data sources such as CVs, websites, and those that may be inferred from those data.

It is important to point out that there might be further limitations of the ethical use of the attributes described in this chapter. Some limitations may originate in cultural aspects. Particularly considering data from CVs, some cultures require while others prohibit the inclusion of attributes such as marital status or age, and therefore their use should consider contextual and local ethical rules.

An outcome of this review is the use of resulting list of attributes as suggestion of attributes to be explored in the next chapters of this thesis. For instance, Chapter 5 focuses on the description of data that will be used in the experiments of the Chapter 7, and such description of data is based on the list of sixty attributes proposed in the conceptual data model for research collaborators.

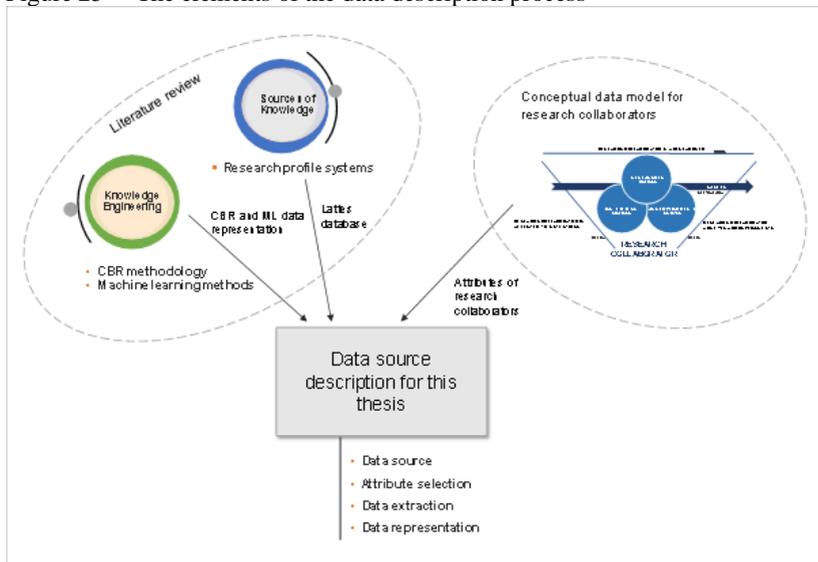


## 5 THE DATA

### 5.1 INTRODUCTION OF THE CHAPTER

This chapter describes the data used in this thesis based on the literature review, described in Chapter 2, and the conceptual data model for research collaborators, presented in Chapter 4. The data description is detailed in order to present the data source, the process of selection the attributes, the data extraction, and how these data will be represented in the *proposed purpose-oriented method*. Figure 25 illustrates the elements used in this process.

Figure 25 – The elements of the data description process



Source: The author, 2017.

The chapter is outlined in five sections including this introduction. Section 5.2 describes the data source; Section 5.3 details the attributes selected from the data source; Section 5.4 presents the data extraction; Section 5.5 represents the data according to the CBR terminology.

## 5.2 THE DATA SOURCE

This thesis explores the use of CV databases as data source, which is a rich source of information on researcher career trajectory (CAÑIBANO; BOZEMAN, 2009; DIETZ et al., 2000). Particularly, this thesis explores data from the Brazilian *Lattes* database (CNPq, 2017), which was previously introduced in section 2.4.7. The *Lattes* is adopted as the data source due to its wealth of information, reliability, completeness, and being freely available for all STI Brazilian institutions, through cooperation agreements with the CNPq. In this study, the access to data from *Lattes* was provided by the Stela Institute ([www.stela.org.br](http://www.stela.org.br)), which apart from cooperation agreements, developed the *Lattes* Platform as a partner of the CNPq, in 1999 (CNPQ, 2017; PACHECO et al., 2006).

## 5.3 THE ATTRIBUTE SELECTION PROCESS

In order to compose a suitable dataset to assess researcher quality, particularly on collaborative purposes, an attribute selection process from the *Lattes* database is performed in a sequence of steps. In the first step, the *Lattes* database is examined in order to search for attributes identified in the *conceptual data model for research collaborators*, for example, *type of accomplishment* (data model) and *type of scientific productions* (*Lattes*). In the second step, *types of scientific productions* are identified in the *Lattes*. After that, in the third step, attributes selected in previous steps are derived into more specific attributes to emphasize collaborative purposes.

Frame 24 lists a set of 10 essential attributes from the *conceptual data model for research collaborators* found in the *Lattes* database. It is important to emphasize that the *Lattes* database provide attributes that characterize researchers in general, however, it does not describe directly the attributes of collaboration. These set of essential attributes will be used in the experiments developed in this thesis.

It is also important to point out that the label of attributes listed in the Frame 24 were adapted from the *Lattes* database to be used exclusively in this thesis. The *Lattes* identifier was included in the set of attributes as personal attribute. It is represented by the Internet address “<http://lattes.cnpq.br/x>”, where “x” is a numeric value of 16 digits that connects the tables of the *Lattes* database.

Frame 24 – List of attributes selected from the *Lattes database* for this thesis

#	Attributes found in the <i>Lattes</i> database	Dimension of the conceptual data model related to the <i>Lattes</i> attribute.
A1	Lattes identifier	Researchers
A2	Institution address	Institutions and affiliations
A3	Fields	Institutions and affiliations
A4	Subfields	Institutions and affiliations
A5	Years since PhD	Career trajectory
A6	Accomplishment type	Accomplishments
A7	Accomplishment year	Accomplishments
A8	Accomplishment co-authors	Accomplishments
A9	Accomplishment sequence coauthoring	Accomplishments
A10	Number of accomplishments	Accomplishments

Source: The author, 2017.

The affiliation attributes characterize researchers' context concerning their institutions and fields. In this type of attribute, the Institution address makes it possible to identify the Brazilian regions and states, and hence, to assess researchers considering their local context, which is advised by the third principle of the Leiden Manifesto, “*Protect excellence in locally relevant research*” (HICKS et al., 2015). Other two relevant attributes are “*fields and subfields*” of knowledge a researcher is affiliated in the institution. The *Lattes* database provides a hierarchical structure to deal with these two attributes called the *CNPQ knowledge areas*, which is composed of “*big area*”, “*area*”, and “*sub-area*”.

The attribute *years since PhD* can be utilized to calculate the productivity of researchers along their career trajectory. The *Lattes* database provides this attribute through the year the researcher received their PhD.

Accomplishments characterize researchers through their scientific achievements. In this study, attributes were selected with the goal of describing characteristics that could imply accomplishments in general, for example, the year and type. Attributes that indirectly characterize the collaborative purpose of the accomplishment were also selected, once that, the *Lattes* database does not provide attributes that characterize research collaboration directly. However, the *Lattes* database provides the number of co-authors, and the sequence of coauthoring of a

production, which allows studies on research assessment to identify collaborative accomplishments.

Thus, in the second step, the researchers' scientific accomplishments are searched in the *Lattes* database. As a result, the research accomplishments found were concerning the bibliographical productions, grant projects, patents, master's dissertation and doctoral thesis advised by them, and their participation as members in committees. Apart from these findings, the *Lattes* database also registers technical, artistic and cultural productions, which due to the scientific purpose of this study, were not considered as relevant. Frame 25 lists accomplishments selected from the *Lattes* database.

Frame 25 – List of accomplishment type identified in the *Lattes database* for this thesis

#	Accomplishment type
a1	Journal articles
a2	Published conference papers
a3	Unpublished conference papers
a4	Books published
a5	Book chapters
a6	Patents
a7	Grants type A (research project)
a8	Grants type B (scholarship)
a9	Grants type C (cooperation)
a10	Grants type D (others)
a11	Master's students - graduated
a12	Master's students - graduated – co-advisor
a13	Doctoral students - graduated concluding
a14	Doctoral students - graduated concluding – co-advisor
a15	Master's students - current
a16	Master's students - current - coadvisor
a17	Doctoral students - current
a18	Doctoral students - current - coadvisor
a19	Membership in master's committees
a20	Membership in doctoral committees

Source: The author, 2017.

Finally, in the third step, in order to emphasize the collaborative purpose of the researcher quality assessment, the first 10 accomplishments listed in Frame 25 were divided in 6 other derived attributes. This process first took into account that co-authorships of scientific papers is a kind of research collaboration (BEAVER; ROSEN, 1978), and that having more than one author (i.e., co-authorships ) can also be used to characterize funded projects as research collaboration (GAZNI; DIDEGAH, 2011; NEWMAN, 2004). Second, by using the number of co-authors, three new types of accomplishments were derived: solo-authored, with one co-author, and with two or more co-authors. Third, by using sequence of coauthoring, such as, first authors, second authors, and third author or other, three more types of accomplishments were derived. At the end of this process, each type of accomplishment identified was considered as an attribute, and a final list of 70 accomplishment type attributes were identified as shown in Frame 26.

Frame 26 – Final list of attributes derived from the accomplishment types, classified as solo or collaborative

#	Attribute	Solo	Collaborative
	<b>Journal articles</b>		
1	number of journal articles solo-authored	x	
2	number of journal articles with one co-author		x
3	number of journal articles with two or more co-authors		x
4	number of journal articles as first author		
5	number of journal articles as second author		x
6	number of journal articles as third author or other		x
	<b>Published conference papers</b>		
7	number of published conference papers solo-authored	x	
8	number of published conference with one co-author		x
9	number of published conference with two or more co-authors		x
10	number of published conference as first author		
11	number of published conference as second author		x
12	number of published conference as third author or other		x

(cont.)

#	Attribute	Solo	Collaborative
13	Unpublished conference papers number of unpublished conference papers solo-authored	x	
14	number of unpublished conference papers with one co-author		x
15	number of unpublished conference papers with two or more co-authors		x
16	number of unpublished conference papers as first author		
17	number of unpublished conference papers as second author		x
18	number of unpublished conference papers as third author or other		x
	<b>Books published</b>		
19	number of books published solo-authored	x	
20	number of books published with one co-author		x
21	number of books published with two or more co-authors		x
22	number of books published as first author		
23	number of books published as second author		x
24	number of books published as third author or other		x
	<b>Book chapters</b>		
25	number of book chapters solo-authored	x	
26	number of book chapters with one co-author		x
27	number of book chapters with two or more co-authors		x
28	number of book chapters as first author		
29	number of book chapters as second author		x
30	number of book chapters as third author or other		x
	<b>Patents</b>		
31	number of patents solo	x	
32	number of patents with one co-author		x
33	number of patents with two or more co-authors		x
34	number of patents as first author		
35	number of patents as second author		x
36	number of patents as third author or other		x

(cont.)

#	Attribute	Solo	Collaborative
	<b>Grants type A (Research project)</b>		
37	number of grants A solo	x	
38	number of grants A with one co-author		x
39	number of grants A with two or more co-authors		x
40	number of grants A as first author		
41	number of grants A as second author		x
42	number of grants A as third author or other		x
	<b>Grants type B (Scholarship)</b>		
43	number of grants type B solo	x	
44	number of grants type B with one co-author		x
45	number of grants type B with two or more co-authors		x
46	number of grants type B as first author		
47	number of grants type B as second author		x
48	number of grants type B as third author or other		x
	<b>Grants type C (Cooperation)</b>		
49	number of grants type C solo	x	
50	number of grants type C with one co-author		x
51	number of grants type C with two or more co-authors		x
52	number of grants type C as first author		
53	number of grants type C as second author		x
54	number of grants type C as third author or other		x
	<b>Grants type D (Others)</b>		
55	number of grants type D solo	x	
56	number of grants type D with one co-author		x
57	number of grants type D with two or more co-authors		x
58	number of grants type D as first author		
59	number of grants type D as second author		x
60	number of grants type D as third author or other		x
61	total number of master's students - graduated		x
62	total number of master's students - graduated – co-advisor		x
63	total number of doctoral students - graduated concluding		x
64	total number of doctoral students - graduated concluding – co-advisor		x

(cont.)

#	Attribute	Solo	Collaborative
65	total number of master's students - current		x
66	total number of master's students - current – co-advisor		x
67	total number of doctoral students - current		x
68	total number of doctoral students - current – co-advisor		x
69	total number of membership in master's committees		x
70	total number of membership in doctoral committees		x

Source: The author, 2017.

In addition to this process, the attributes were examined considering firstly “*co-authorship*” in publications and research projects as research collaboration (e.g., BEAVER; ROSEN, 1978; GAZNI; DIDEGAH, 2011; NEWMAN, 2004). After that, these attributes were classified as solo or collaborative. The 60 first attributes, with the exception of the attribute “number of accomplishment as first author”, were classified as solo or collaborative. However, the “*number of accomplishment as first author*” was not discernable as solo or collaborative, because if researchers have solo accomplishments, then, in this accomplishment the researcher will be the first author. On the other hand, if researchers have accomplishments with one co-author or more, they may be the first author or second, etc.

The attributes number 61 to 68 concerns mentoring activities, when a professor collaborates with a junior scientist, a post-doctoral researcher, or a graduate student (e.g., BOZEMAN; CORLEY, 2004). This kind of activity is also associated with collaboration in invisible colleges (PRICE; BEAVER, 1966). Furthermore, Tuesta et al. (2015) comments about the positive impact of collaboration with the advisor to the PhD graduate publications. This study uses data from the Brazilian *Lattes* database (lattes.cnpq.br), in which the authors were able to identify attributes that characterize the advisor-advisee relationship. In this thesis, I identified in the *Lattes* database, the total number Master and Doctor mentoring activities, and classified such attributes as collaborative.

The last two attributes are associated to collaboration between professors during examination committees for candidates to Masters or Doctor degrees. Neto and Da Costa, (2016) address this issue by using data from the Brazilian *Lattes* database (lattes.cnpq.br), in the area of Accounting Sciences in Brazil. The authors use the attribute total number of membership in master's committees in their analysis, and highlights

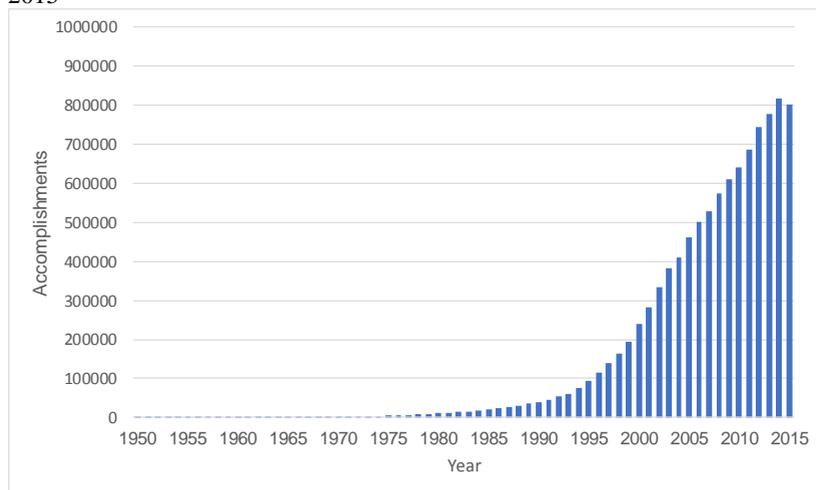
that this kind of collaboration is established in a context apart from co-authorship in articles, books, or another publication.

#### 5.4 THE DATA EXTRACTION PROCESS

The data for this study were extracted from the Lattes database in July 2016. A dataset with about 4 million curriculums was provided for this study by the Stela Institute ([www.stela.org.br](http://www.stela.org.br)), which downloaded it from the CNPq, through agreements of cooperation. The data extraction was limited to PhD researchers. It was also limited to accomplishments achieved after the conclusion of their PhD. Because the data from the year 2016 was incomplete, accomplishments of this year was discarded. After that, for each accomplishment listed in Frame 26, the researcher productivity (P) was calculated as the total of accomplishments achieved by researchers a year, along their career trajectory.

At the end of the data extraction, about 208,150 CVs were obtained, with accomplishments registered from 1950 to 2015. Figure 26 illustrates the evolution of research accomplishments from 1950 to 2015.

Figure 26 – Evolution of the Brazilian researcher accomplishments from 1950-2015



Source: The author, 2017. Data obtained from the data extraction process for this thesis, in July 2016, from the data set provided by the Stela Institute <sup>5</sup>, through agreements of cooperation with the CNPq.

<sup>5</sup> Stela Institute: [www.stela.org.br](http://www.stela.org.br)

In addition to researcher productivity (P), in the end of the process, research affiliations to institutions and fields, and the year in which researchers concluded their PhD, were added to the final dataset.

## 5.5 THE DATA REPRESENTATION PROCESS

As described in Chapter 3, a purpose-oriented method was proposed to assess research quality assessment. The proposed approach is designed by using the CBR methodology and Machine Learning methods (ML). The CBR methodology represents cases through any ML representational formalism, such as feature vectors (i.e., tables), objects, textual, and semantic representations (e.g., ontologies, XML, OWL), among others (EL-SAPPAGH; ELMOGY, 2015). In this thesis, I adopt the simplest CBR representation, that is, a feature-value pairs data structure (e.g., RICHTER; WEBER, 2013). In future works I intend to explore other representational formalisms, for example, ontologies.

In the feature-value pairs representation, a case base (CB) (i.e., a collection of cases) is organized as an attribute-value vector (i.e., a table of rows and columns), where each row describes a case (c), and each column represents one of the attributes (a). The value of each cell of the table (i.e., row x column) represents the value of each attribute in the case (v). In addition, the last column of the table is the class that distinguishes each case. Table 4 illustrates the structure of a case base (CB).

Table 4 – The case base (CB) represented by a feature-value pairs data structure

Cases	Attribute 1 (a <sub>1</sub> )	Attribute 2 (a <sub>2</sub> )	...	Attribute n (a <sub>n</sub> )	Class
Case 1 (c <sub>1</sub> )	v <sub>1,1</sub>	v <sub>1,2</sub>	...	v <sub>1,m</sub>	class 1
Case 2 (c <sub>2</sub> )	v <sub>2,1</sub>	v <sub>2,2</sub>	...	v <sub>2,m</sub>	class 2
Case 3 (c <sub>3</sub> )	v <sub>3,1</sub>	v <sub>3,2</sub>	...	v <sub>3,m</sub>	class 2
...	...	...	...	...	class 1
Case m (c <sub>m</sub> )	v <sub>n,1</sub>	v <sub>n,2</sub>	...	v <sub>n,m</sub>	class 1

Source: Adapted of Richter and Weber (2013).

Considering that cases are experiences, in the proposed *purpose-oriented method*, a case base of CVs refers to the scholarly experiences of researchers achieved along their career trajectory. This collection of cases is represented by *feature-value pairs*, as described above.

Table 5 – The case base of CVs represented by a feature-value pairs data structure

<b>Researchers</b>	<b>Article (a<sub>1</sub>)</b>	<b>Book (a<sub>2</sub>)</b>	<b>Patent (a<sub>3</sub>)</b>	<b>...</b>	<b>Grant (a<sub>n</sub>)</b>	<b>Class</b>
Researcher 1 (r <sub>1</sub> )	10	3	1	...	3	fit
Researcher 2 (r <sub>2</sub> )	5	3	0	...	1	fit
Researcher 3 (r <sub>3</sub> )	40	9	2	...	10	unfit
...	...	...	...	...	...	unfit
Researcher m (r <sub>m</sub> )	15	5	1	...	2	fit

Source: The author, 2017.

Each column of the table describes a researcher (r) and each row describes one type of accomplishment (a). Each cell value (v) represents the research productivity (P) in each type of accomplishments. At the end of the table, a class column distinguishes each researcher as *fit* (i.e., positive instance) or *unfit* (i.e., negative instance) for a purpose of the assessment. Table 5 illustrates the case base of CVs represented by a *feature-value pairs* data structure.

## 5.6 CONCLUDING REMARKS

Chapter 5 described the data that will be used in the experiments of this thesis, in order to demonstrate the the proposed *purpose-oriented method*. More precisely, this chapter described the data source; detailed the essential attributes selected from the data source to demonstrate the proposed method; presented the data extraction process; and at the end, represented the set of data according to the CBR terminology.

Thus, firstly the Brazilian *Lattes* database (CNPq, 2017) was chosen as a suitable data source to this study, due to its wealth of information, reliability, completeness, and being freely available for all STI Brazilian institutions (LANE, 2010; PERLIN et al., 2017).

After that, the *Lattes* database was examined in order to search for the attributes identified in the *conceptual data model for research collaborators*. So, considering that the purpose of *Lattes* database is to provide attributes that characterize researchers in general, and not to describe directly attributes of collaboration, from the 60 attributes suggested in Chapter 4, 10 essential attributes was found, which are sufficient to develop the experiments. However, in future works the universe of attributes suggested by the data model could be better explored.

Then, each type of accomplishment identified in the *Lattes* database was considered as an attribute, and each one of attributes were categorized as “*solo*” or “*collaborative*” purposes, based on literature. As a result, a final list of 70 attributes were delivered as shown in Frame 26, emphasizing the collaborative purpose of the researcher quality assessment.

Having consolidated the set of attributes, the next step was the data extraction process, which resulted in 208,150 CVs, with accomplishments registered from 1950 to 2015. Concluding this process of data description, the simplest CBR representation, that is, a *feature-value pairs data structure* (e.g., RICHTER; WEBER, 2013) was adopted to represent the data in the proposed method.

In the next chapter, the *purpose-oriented method to assess researcher quality* will be introduced, and specific requirements of data will be specified in more details.

## 6 THE PURPOSE-ORIENTED METHOD

### 6.1 INTRODUCTION OF THE CHAPTER

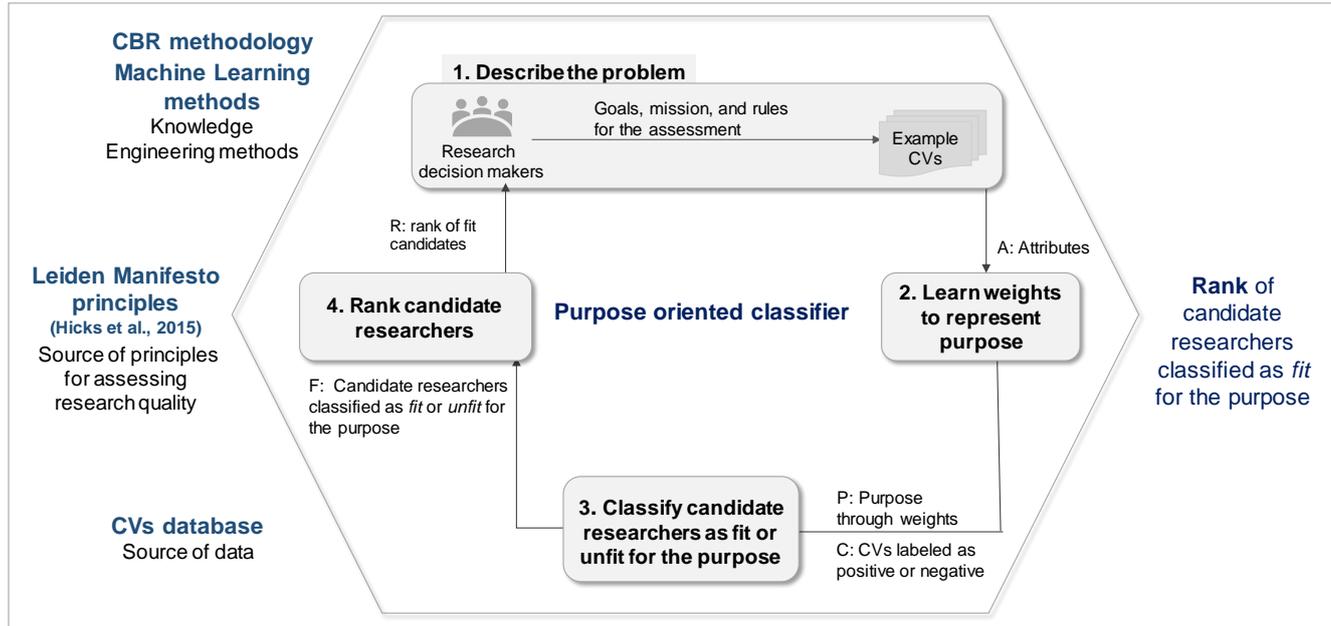
Chapter 6 introduces the *purpose-oriented method* to answer the main research question of this thesis, “*How to assess researcher quality for collaborative purposes?*”. After the contextualization of the problem, the investigation of the knowledge about this problem in the literature, and the description and representation of the data, I outline the implementation of the proposed method.

Considering researcher quality assessment is necessary in selection processes such as recruitment, promotion, and grant awarding decisions (GARFIELD; MALIN, 1968; HAUSTEIN; LARIVIÈRE, 2015; LANE, 2010; LANE et al., 2015), this thesis proposes the *purpose-oriented method to assess researcher quality*. The method taken into account that experiences achieve by researchers along their career trajectories are mapped into their CVs, and are represented by their accomplishments (i.e., research productions). Thus, by analogy with the CBR principle, in which a new problem is solved by comparing previous similar experiences (e.g., AAMODT; PLAZA, 1994), the *purpose-oriented method* assesses the similarity between new candidate researchers with successful researchers, through their accomplishments.

Figure 27 presents a general view of the *purpose-oriented method*, whose core is a classifier denominated *purpose-oriented classifier*, which is able to classify a researcher as fit or unfit for the purpose the assessment.

The approach initiates by defining fundamental inputs for the *purpose-oriented classifier*. Firstly, the proposed method relies on the Leiden Manifesto for research metrics (HICKS et al., 2015) as source of principles for assessing research quality. As previously mentioned in Section 2.2.6, the Leiden Manifesto gathers 10 principles, which have been widely recommended by the bibliometrics community (e.g., BORNMANN; HAUNSCHILD, 2016; COOMBS; PETERS, 2017; OECD, 2016).

Knowledge Engineering methods provide to the purpose-oriented classifier, the CBR methodology and ML methods. These KE methods will engineer computational solutions supporting the classifier.

Figure 27 – The *purpose-oriented method* to assess researcher quality

Source: The author, 2017.

The CBR methodology, as described in Section 2.5.3, is used to implement the approach, which relies more precisely on the CBR principle, as previously mentioned. In addition to this principle, CBR provides the approach a cycle of development in four steps: *Retrieve*, *Reuse*, *Revise*, and *Retain*. By taking into account that a case represents experience, the cycle incorporates “*what to do*” in order to find useful experiences (RICHTER; WEBER, 2013). The CBR methodology is a mature methodology with a vast availability of methods, which may be combined in each step of its cycle (AHMED; BEGUM; FUNK, 2012; WATSON, 1999). Furthermore, the CBR methodology has been adopted for different application areas, and for both analytical or synthetical tasks. For example, the synthetical task of planning a business process (AGORGIANITIS et al., 2016), and the analytical task of investigating time series for predictions (GUNDERSEN, 2014; KURBALIJA, 2009).

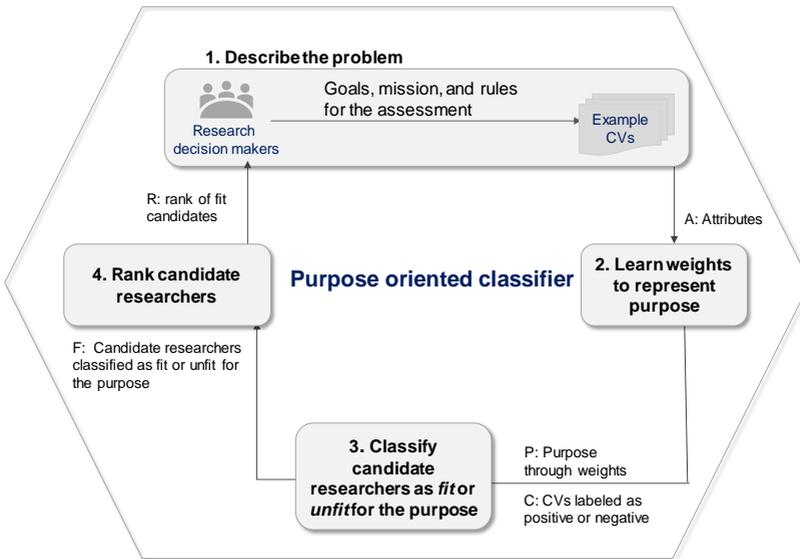
Machine Learning methods entail the implementation of computational solutions that are applied in the attributes described on Section 5.3. It concerns training algorithms that learn knowledge from data, as described on Section 2.5.4. The training algorithms produce a classifier, which is a data structure that can classify, for example, a researcher CV as fit or unfit for a purpose.

Lastly, the *purpose-oriented classifier* receives as input, data from the Brazilian *Lattes* database (lattes.cnpq.br), which was chosen as data source. It allows this study to explore contents from CVs that include local references, but that may be universally applicable. This thesis had access to more than 4 million curriculums from one of the cleanest and highest quality researcher databases (LANE, 2010). Such quality data supplies the purpose-oriented classifier the opportunity of applying the proposed method considering the Leiden Manifesto principles (HICKS et al., 2015).

The purpose-oriented classifier will be described in the next section.

## 6.2 THE PURPOSE-ORIENTED CLASSIFIER

In this section, the CBR implementation for the *purpose-oriented classifier* is described and applied through ML methods to the data from the Brazilian *Lattes* database (lattes.cnpq.br). The purpose-oriented classifier is an implementation of similarity heuristics, and it is illustrated in Figure 28.

Figure 28 – The *purpose-oriented classifier*

Source: The author, 2017.

Firstly, before describing the proposed classifier, it is necessary to define three intertwined terms, which are, purpose, quality and context. The term *purpose* means intention, aim, goal, and propose (e.g., [www.etymonline.com](http://www.etymonline.com)). According to Juran and Godfrey (1999), *quality is fitness for purpose*, because quality depends on user perspectives, needs and priorities, and vary across user-groups. Thus, the *purpose-oriented classifier* adopts the definition of Juran and Godfrey (1999), that quality is fitness for purpose, and that assessing the quality of researchers depends on the decision makers' quality requirements for the target assessment. The third term, *context*, is defined by Abowd et al.(1999) as any information that can be used to characterize the situation of an entity considered relevant to the interaction between a user and an application. For instance, ML methods characterize the context of a domain area by learning from its attributes (HALL, 1999; WITTEN; FRANK, 2005).

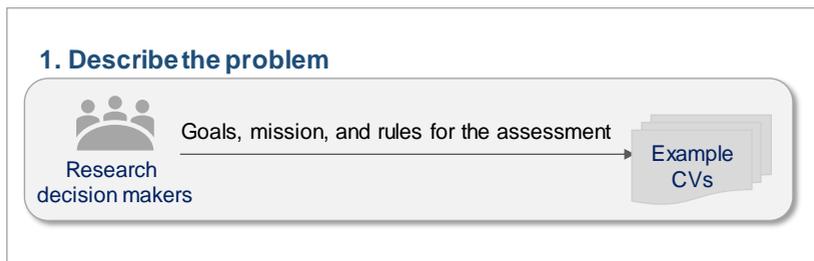
The *purpose-oriented classifier* is tackled in four stages. The goal of *Stage 1* is to describe the assessment through the intended quality requirements of decision makers. After that, in *Stage 2*, the purpose of the assessment is characterized through weights by applying a pre-processing

ML approach. Then, in *Stage 3*, the purpose through weights is used in a CBR implementation, in order to classify researchers as fit or unfit for the purpose. Closing the cycle, in *Stage 4* researchers are ranked in a list that is delivered to decision makers, who may reuse and adapt it in new research assessment cycle. These four stages will be detailed in the next subsections.

### 6.2.1 Stage 1: Describing the problem

*Stage 1*, as illustrated in Figure 29, is receiving the intended quality requirements (i.e., goals, mission, and rules) for the assessment process from decision makers.

Figure 29 – Stage 1: Describing the target problem of the research assessment



Source: The author, 2017.

For example, suppose a job opens at UFSC for a professor on knowledge engineering, to work for the department of knowledge engineering and management. The objective of this assessment would be selecting candidates who have a large experience and many accomplishments in data science. Another example is the case of funding agencies that intend to foment a project that needs: researchers able to work in an interdisciplinary research collaboration, to collaborate with ten people or more, and to produce valuable results.

There are two ways of gathering these quality requirements from decision makers. The requirements can be provided in a direct and objective way. For example, candidates with a PhD in a specific area. This direct way can be represented by rules or by numerical models that simply deliver an exact result, such as the answer “yes” or “no”. However, this objective way is usually not available.

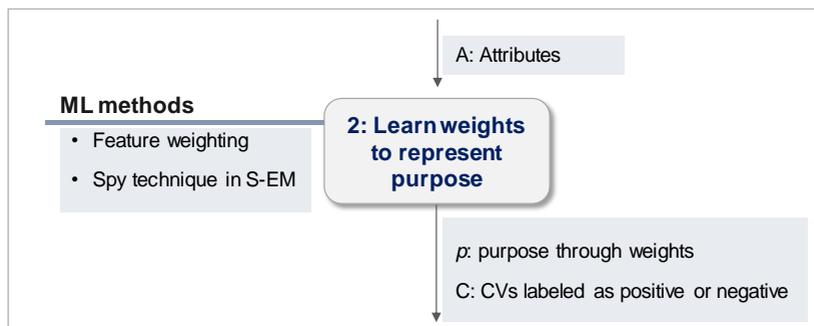
In contrast to the direct way, the quality requirements may be provided by decision makers in a subjective way, by exemplars. For

instance, when the users say, “*We want a professor with a PhD from a university with the same level of quality as ours*”. This example is very subjective, because it makes categorizing universities necessary to obtain those with the same level of quality desired. Other subjective example is when decision makers exemplify: “*we want such researchers, but we also do not want these others*”. Another and even more complex situation is when decision makers provide only good examples. Despite this complexity, in all of these situations, if such exemplars of researchers are registered in CVs, ML methods could be applied to automatically characterize the context of such exemplar CVs, and translate the decision makers quality requirements that are imbedded in the exemplar CVs into knowledge about the problem.

### 6.2.2 Stage 2: Learning weights to represent purpose

*Stage 2* is a pre-processing step that receives as parameter, the attributes of the set of example CVs, and evaluates the relative relevance of each attribute through ML methods. There are two results, the purpose of the assessment learned through weights ( $p$ ), and the set of example CVs labeled as positive or negative ( $C$ ). Figure 30 illustrates Stage 2.

Figure 30 – Stage 2: Learning purpose through weights



Source: The author, 2017.

The Leiden Manifesto (HICKS at al., 2015) recommends in its second principle that performance should be measured taking into account the research missions of the institution, group or researcher. However, one of the difficulties of following this principle, is that the purpose (i.e., mission) of the assessment is often unclear, and too general (COOMBS; PETERS, 2017). The other, according to the authors of the

Manifesto themselves, is that, no single evaluation model applies to all contexts (HICKS at al., 2015).

In a work published while developing this thesis, I investigated how considering purpose influences the accuracy of the assessment (DUARTE; WEBER; PACHECO, 2016a). In such study, I applied *feature weighting* (e.g., WETTSCHERECK; AHA; MOHRI, 1997) to characterize the purpose of the assessment, and demonstrated the hypothesis:

*H2.1: A purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method.*

Feature weighting emphasizes the relevance of each attribute, by assigning weights to each one, in which different degrees of relevance are given, and none of the attributes is discarded (WETTSCHERECK; AHA; MOHRI, 1997). However, because *feature weighting* is a supervised method, it requires positive and negative instances to be learned (CARBONELL; MICHALSKI; MITCHELL, 1983).

In the work aforementioned, I used the same representational structure described on Section 5.5, and data instances were labeled by humans through commonsense knowledge, as *fit* (i.e., positive instance) or *unfit* (i.e., negative instance) for the purpose, in order to train the classifier. This work demonstrated the viability of purpose-oriented methods to assess researcher quality, and allowed me to continue this investigation exploring other situations. This work will be better detailed in Chapter 7 – Experiment I.

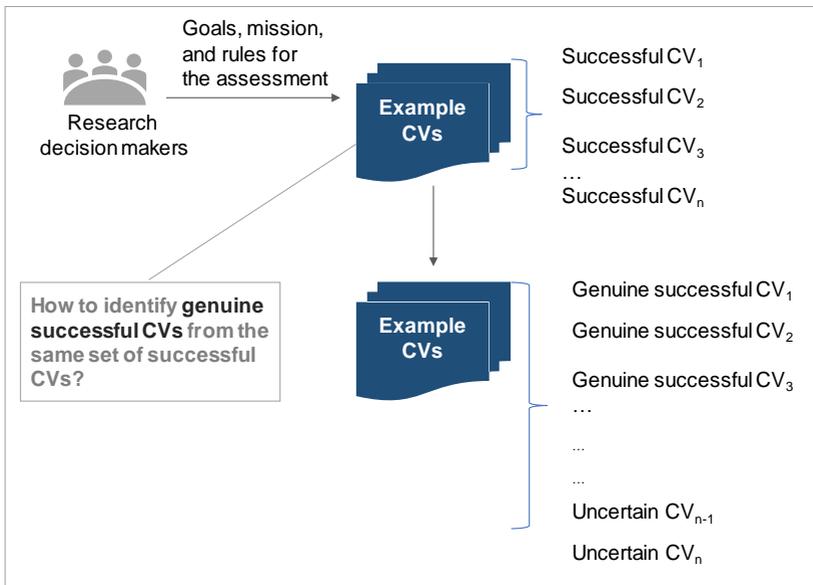
Considering findings of this previous work, a different example was explored, in which the data contains only positive instances, i.e., the quality requirements of decision makers contain only successful CVs. This problem is addressed by Liu et al. (2002), who propose a *partially supervised* classification approach called the *Spy technique in S-EM*, which was previously described in Section 2.5.4.3.

In summary, the approach of Liu et al. (2002) simulates an investigation on the behavior of unknown positive documents, in an attempt to identify those with characteristics that differ them from the others. The study of Gunawardena et al. (2013) applies this technique to a CBR problem, attempting to identify cases that are well aligned (i.e., positive instances) versus cases that are poorly aligned (i.e., negative instances).

### 6.2.2.1 Applying the Spy technique in S-EM to the purpose-oriented method

The problem faced here is that the set of quality requirements (i.e., example CVs) given by decision makers contains only successful CVs, and to evaluate the relative relevance of each attribute, and thus, learn the purpose of the assessment through weights, there should be successful and unsuccessful CVs. More specifically, the problem is the lack of negative instances to learn weights. Thus, a suitable approach should be to identify genuine successful CVs and uncertain successful CVs, from the set of example CVs. Figure 31 illustrates this problem.

Figure 31 – The problem of finding genuine successful CVs.



Source: The author, 2017.

In order to label the example CVs as positive or negative, I applied the *Spy technique in S-EM* (LIU et al., 2002). The approach begins by considering two sets of CVs: (i) the set of example CVs, contains instances labeled as positive, and thus, it is called of positive *set* ( $P$ ), which is defined as  $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_n)$ ; and (ii) the set of candidate CVs contains unlabeled instances, and so, it is called of unlabeled *set* ( $U$ ), which is defined as  $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ . The *sets*  $P$  and  $U$  have the same

structure, and they are organized as a *feature-value pairs data structure* (i.e., a table of rows and columns), where each row describes a researcher ( $r_n \in \mathbf{R}$ , with  $\mathbf{R} = (r_1, \dots, r_n)$ ), and each column represents one of their attributes ( $a_m \in \mathbf{A}$ , with  $\mathbf{A} = (a_{11}, \dots, a_{mk})$ ). The value of each cell of the table (i.e., row  $\times$  column) represents the absolute value of productivity of a researcher in each type of attribute. The list of attributes was described in Section 5.3, and Table 6 illustrates this data structure.

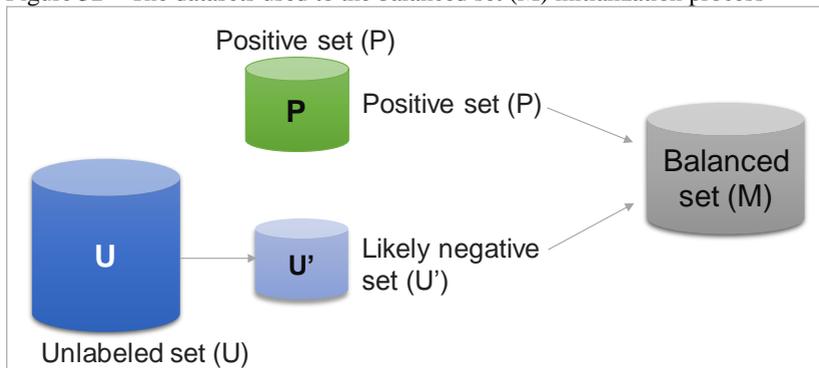
Table 6 – The *positive set (P)* represented by a *feature-value pairs data structure*

Researcher	(a <sub>1</sub> )	(a <sub>2</sub> )	(a <sub>3</sub> )	...	(a <sub>n</sub> )	Class
r <sub>1</sub>	10	3	1	...	3	P
r <sub>2</sub>	5	3	0	...	1	P
r <sub>3</sub>	40	9	2	...	10	P
...	...	...	...	...	...	P
r <sub>n</sub>	15	5	1	...	2	P

Source: The author, 2017.

The next step is the initialization of a balanced set of both positive and negative instances. Despite the negative instances being unknown, the approach of Liu et al. (2002), takes into account that is possible to identify some very likely negative instances from the *unlabeled set (U)*. To this end, I create a subset of *likely negative instances (U')*, which contains instances from the *set (U)* that are completely opposite to the *positive instances (P)*. Figure 32 illustrates the three data sets used to create the *balanced set (M)* to perform the *Spy technique in S-EM* (LIU et al., 2002).

Figure 32 – The datasets used to the balanced set (M) initialization process



Source: The author, 2017.

For selecting the set of likely *negative instances* ( $U'$ ), a strategy is proposed based on the assumption that *likely negative instances* are opposite to the *positive instances*. Thus, taking into account that in the *set*( $P$ ) there are attributes with *higher productivity* ( $H$ ) and also attributes with *lower productivity* ( $L$ ), the intention is firstly to identify these two subsets of attributes.

For this, given the *set* ( $P$ ) and its attributes  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)$ , the sum of each attribute is calculated considering the entire set  $\mathbf{A}$  (i.e.,  $\text{sum}(\mathbf{a}_i) = \sum \mathbf{a}_i, i=1\dots n$ ). After that, each attribute  $\mathbf{a}_i$  is analyzed, and a threshold ( $t$ ) is considered to decide whether an attribute have *higher productivity* ( $H$ ) (i.e.,  $\text{sum}(\mathbf{a}_i) > t$ ) or *lower productivity* ( $L$ ) (i.e.,  $\text{sum}(\mathbf{a}_i) \leq t$ ).

After that, two filters can be created to search for the likely negative instances in the *set* ( $U$ ). One has the goal of finding instances in the *set* ( $U$ ) using the subset of attributes with higher productivity  $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_m)$ , where opposite instances to the *set* ( $P$ ) are searched.

$$U' \leftarrow (h_1 = 0 \text{ and } h_2 = 0 \text{ and } h_3 = 0 \dots \text{and } h_k = 0)$$

The other filter, intends to find instances in the *set* ( $U$ ) using the subset of attributes with *lower productivity* ( $L$ ).

$$U' \leftarrow (l_1 = 0 \text{ and } l_2 = 0 \text{ and } l_3 = 0 \dots \text{and } l_k = 0)$$

This strategy was developed based on a series of experiments performed using data of two application scenarios, which lead me to acceptable results. These experiments will be detailed in Chapter 7 – Experiments, and in future works, I intend to study the sensibility of such strategy, in order to propose those that could be more generalizable.

Finally, a balanced *set* ( $M$ ), of both *positive* and *negative* instances, is composed with instances from the *set* ( $P$ ) and the *set* ( $U'$ ). Having initialized the balanced *set* ( $M$ ), the *Spy technique in S-EM* (LIU et al., 2002) is applied in five steps, as illustrated in Figure 33.

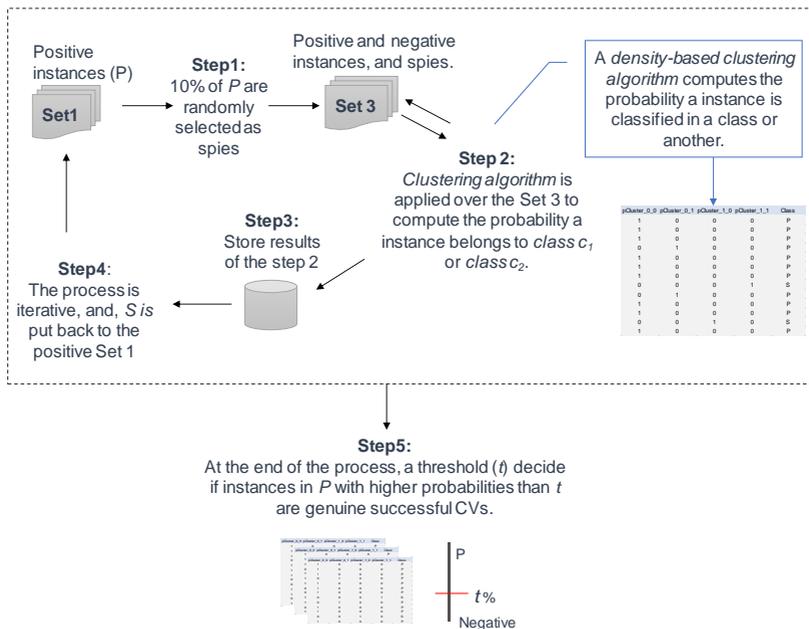
The process is iterative, and to each cycle in Step1, 10% of *positive instances* ( $P$ ) are selected as “*spies*” to be included in the *balanced set* ( $M$ ). The final *set* ( $M$ ) contains *positive*, *spies*, and likely *negative instances*. The *positive instances* are the class  $c_1$ , and the remaining instances are the class  $c_2$ .

In *Step 2*, a *density-based clustering algorithm* computes the probability an instance is classified in classes  $c_1$  or  $c_2$ . To this end, an

extension of the *EM algorithm* (DEMPSTER; LAIRD; RUBIN, 1977) called the *cluster membership filter*, which is implemented in the Weka tools (<http://www.cs.waikato.ac.nz/ml/weka/>), is used. Initially, each positive instance  $p_i \in P$  is assigned the probability  $\Pr[c_1, p_i] = 1$  and  $\Pr[c_2, p_i] = 0$ .

In *Step 3*, the results are stored to compute the final results, and in *Step 4*, the spies are put back in the *positive set* ( $P$ ).

Figure 33 – The Spy technique in S-EM (LIU et al., 2002) applied to the purpose-oriented method



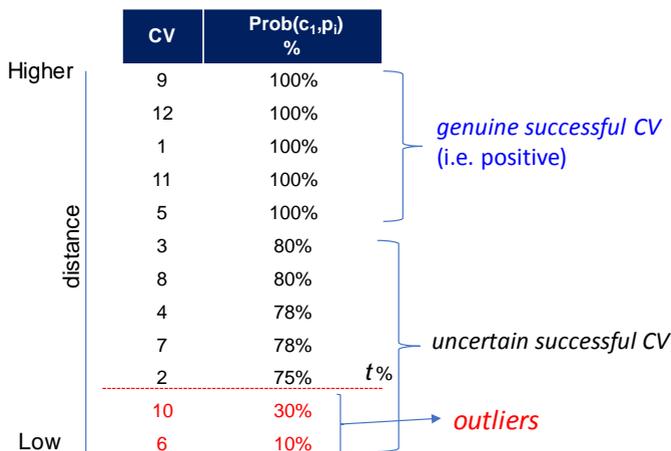
Source: The author, 2017.

In *Step 5*, in the end of the iterative process, as illustrated in Figure 34, the probability of a CV belong to class  $c_1$  is sorted from the highest to the lowest, showing the alignment between the CVs.

Then, CVs with the highest probability (i.e., 100%) are considered genuine successful CVs, and the remaining ones are considered uncertain successful CVs. At last, from the uncertain successful CVs, a threshold ( $t$ ) is used to limit the outliers (i.e., the most

distant CVs from the genuine successful CVs). Liu et al, (2002) used a threshold ( $t$ ) of 15% to classify documents with probability lower than  $t$  as *negative* ( $N$ ). However, in the *purpose-oriented method* the intention is to define a threshold ( $t$ ) not in relation to a percentage, but with the goal to limit the outliers identified.

Figure 34 – Example of the *alignment of case CVs by ranking*



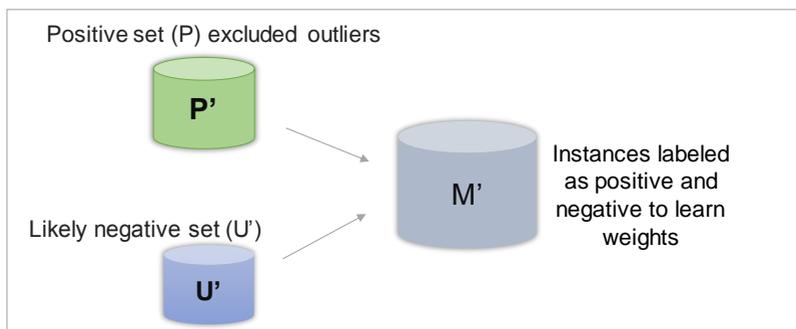
Source: The author, 2017.

After categorizing the exemplar CVs in genuine successful CVs, uncertain successful CVs, and outliers, the next step is to characterize the context of set of exemplar CVs. ML methods, as previously mentioned, characterize the context of a domain area by learning from the relevance of its attributes (ABOWD et al., 1999; HALL, 1999; ROBNIK-ŠIKONJA; KONONENKO, 2003). Thus, the goal of this step is to apply *feature weighting* algorithms to rank the relevance of attributes from the example CVs, by assigning weights to them. Consequently, the purpose of this group of CVs will be reflected through the relevance of such attributes.

For applying feature weighting algorithms, a set  $M' = (m'_1, \dots, m'_n)$  was composed of two sets: (i) the *positive set* ( $P$ ) excluded those instances that are outliers, which I called  $P' = (p'_1, \dots, p'_n)$ ; and (ii) the *set of likely negative instances*  $U' = (u'_1, \dots, u'_n)$ . Instances from the set ( $P'$ ) were labeled as positive and instances from the set ( $U'$ ) where

labeled as negative. Figure 35 illustrates the set of positive and negative instances to apply *feature weighting* algorithms.

Figure 35 – The dataset of *positive* and *negative* instances to apply *feature weighting* algorithms



Source: The author, 2017.

Examples of feature weighting algorithms used in this step, are: Correlation Based Feature Selection (CFS) (HALL, 1999); Information Gain (IG) (QUINLAN, 1992); Symmetrical Uncertainty (PRESS et al., 1988), and ReliefF (KIRA; RENDELL, 1992; KNONENKO; ROBNIK-SIKONJA; POMPE, 1996).

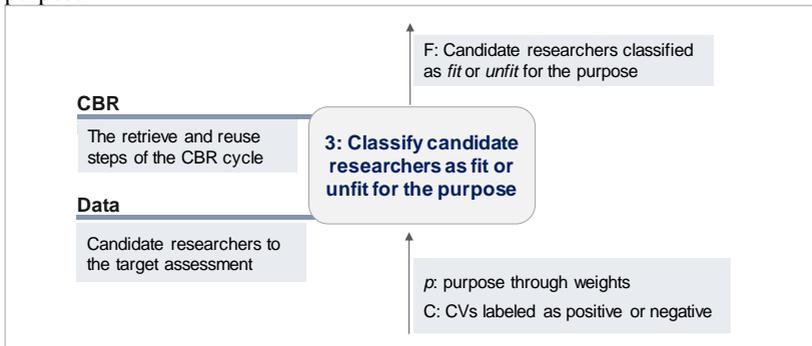
### 6.2.3 Stage 3: Classifying candidate researchers as *fit* or *unfit* for the purpose

In *Stage 3* the second research question, “*How to assess researcher quality?*” is effectively answered. The proposed approach, illustrated in Figure 36, is a CBR implementation, in which the first two steps concerns the design CBR steps; the third step implements the retrieve step of the CBR cycle; and the fourth step implements the reuse step of the CBR cycle:

1. The CBR representation is defined;
2. The datasets are populated;
3. The similarity score between a new case (i.e., a candidate researcher) and each case (i.e., the example CV) is calculated, and a set of cases sufficiently similar to the new case are returned.

4. the most similar case ( $c_j$ ) to a new case ( $q_i$ ) is chosen to generate the classification of a candidate researcher as fit or unfit for the purpose of the assessment.

Figure 36 – Stage 3: Classifying candidate researchers as *fit* or *unfit* for the purpose



Source: The author, 2017.

The *first task* begins by describing the inputs of *stage 3* in a CBR representation. These inputs are: the set of CVs labeled as *positive* or *negative*; the purpose represented by the weights; and the set of unlabeled CVs of candidate researchers to the target assessment.

In this thesis, as previously described in Section 5.5, the simplest CBR representation is adopted, that is, a *feature-value pairs data structure* (e.g., RICHTER; WEBER, 2013), as illustrated in Table 7.

The set of CVs labeled as *positive* or *negative* is represented by *cases* ( $C$ ) defined as  $C = (c_1, \dots, c_m)$ , where  $m$  represents the total number of cases under consideration. The set of *candidate CVs* is represented by *new cases* ( $Q$ ) defined as  $Q = (q_1, \dots, q_n)$ , where  $n$  represents the total number of new cases under consideration. The *set* ( $Q$ ) contains unlabeled instances.

Table 7 – The CBR representation for the *purpose-oriented classifier*

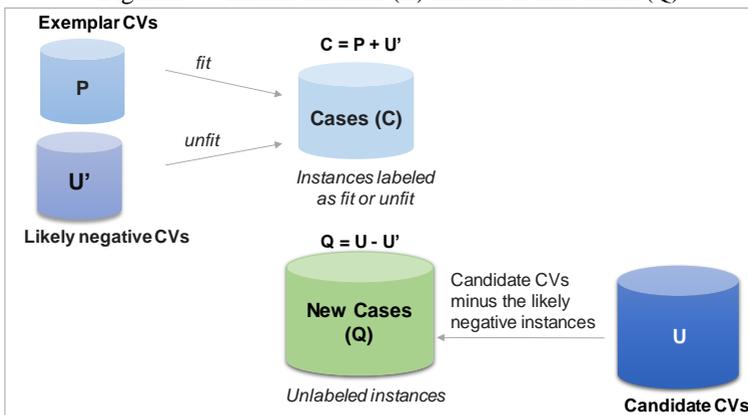
Cases	Attribute 1 ( $a_1$ )	Attribute 2 ( $a_2$ )	...	Attribute n ( $a_n$ )	Class
Case 1 ( $c_1$ )	$v_{1,1}$	$v_{1,2}$	...	$v_{1,m}$	Positive
Case 2 ( $c_2$ )	$v_{2,1}$	$v_{2,2}$	...	$v_{2,m}$	Negative
Case 3 ( $c_3$ )	$v_{3,1}$	$v_{3,2}$	...	$v_{3,m}$	Negative
...	...	...	...	...	Positive
Case m ( $c_m$ )	$v_{n,1}$	$v_{n,2}$	...	$v_{n,m}$	Negative

Source: Adapted of Richter and Weber (2013).

The cases (C) and new cases (Q) are represented through the same data structure described above, where each row lists a researcher ( $r_n$ )  $\in R$ , with  $R = (r_1, \dots, r_n)$ , and each column represents one of their attributes ( $a_m$ )  $\in A$ , with  $A = (a_{11}, \dots, a_{mk})$ . The value of each cell of the table (i.e., row x column) represents the absolute value of productivity of a researcher in each type of attribute selected based on the raw CV data, as described in Section 5.3. In addition, the purpose represented by the weights is defined by the set  $W = (w_1, \dots, w_m)$ , where each  $w_i$  is a value between [0,1].

The *second task* is populating the datasets with data from the Stage 2. Thus, the set of cases (C) is composed of the positive set (P) and the set of likely negative instances (U'). The set of new cases (Q) is composed of the unlabeled set (U) minus the set of likely negative instances U'. Figure 37 illustrates the final datasets C and Q that were used in Stage 3.

Figure 37 – The set of cases (C) and set of new cases (Q)

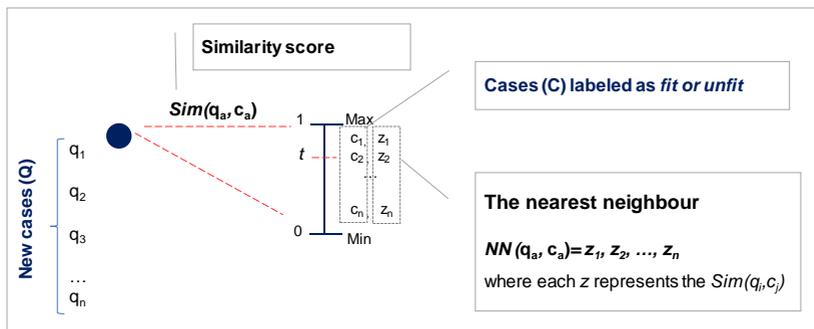


Source: The author, 2017.

The *third task* is calculating the similarity score between each *new case* ( $q_i$ ) and *all cases* ( $c_1, \dots, c_m$ ), that is, the global similarity between them. As a general rule, a global similarity score is normalized in a range of zero to one, where zero is given when the values are totally dissimilar, and one when the values are totally similar. When there are two values that are different from each other but that may be considered proportionally similar, a value between zero and one is assigned. The *global similarity measure* was previously presented in subsection 2.5.3.5, The local-global principle for similarity measures.

The third task results in a set of similar cases to a new case, which is called the nearest neighbour  $NN(q_a, c_a)$ , as illustrated in Figure 38. The nearest neighbour is a relation that is not usually uniquely defined, for instance, there may be several equally similar cases to each new case.

Figure 38 – The neighbour ( $NN_k$ ) between a new case (Q) and cases (C)



Source: The author, 2017.

The fourth task aims at identifying the most suitable solution that represents the entire set of nearest neighbours, and classifying a candidate researcher as fit or unfit for the purpose of the assessment. In order to do so, I used an ad-hoc strategy based on analyzing the precision of the similarity score ( $Z_i$ ), which is depicted in the following sequence of steps:

1. The set of *nearest neighbours* to a *new case* ( $q$ ) is listed in a *neighbour ranking*, from the highest to lowest. This rank is characterized by the cases (C) labeled as positive or negative, and their respective similarity score  $z_i = Sim(q_a, c_a)$ , as illustrated in Table 8.

Table 8 – The rank of similar cases to each new case

Cases (C)	Class	Similarity Score $Z_i = Sim(q_i, c_j)$
Case 1 ( $c_1$ )	positive	0.9974
Case 2 ( $c_2$ )	positive	0.9870
Case 3 ( $c_3$ )	negative	0.9765
...	...	...
Case m ( $c_m$ )	negative	0.8856

Source: The author, 2017.

2. The set of neighbour ranking is analyzed, and if many redundant negative cases are found, that is, cases with equivalent values of  $z_i$ , I adopt a strategy to reduce the amount of negative cases. This strategy is needed if the set of cases has few positive cases and many negative cases:
  - i. Given the set of neighbour ranking, I selected a sample of 10% of cases and analyzed those of representative behavior;
  - ii. The similarity score ( $Z_i$ ) with lowest value is defined as a threshold ( $t$ );
  - iii. The set of neighbour ranking is limited to those with  $(Z_i) > t$ ;
  - iv. Finally, the similarity score ( $Z_i$ ) with a precision of four decimal places is used to remove the negative cases.
3. In order to determine if a new case is *fit* or *unfit* for the purpose of the assessment, a rule is applied considering the top three nearest neighbours, their respective classes (positive or negative) and similarity scores ( $Z_i$ ). This rule is algorithmically presented in the Pseudocode-1.

**Pseudocode-1:** Classify a new case as *fit* or *unfit*

**Begin**

**Input:** Case  $C = (c_1, \dots, c_m)$

Class  $C = [\text{positive, negative}]$

$NN = (z_1, \dots, z_k)$ , where  $z_i = \text{Sim}(q_i, c_j)$

$Q = (q_1, \dots, q_n)$

Class  $Q = [\text{fit, unfit}]$

**Output:** The new case ( $Q$ ) classified as *fit* or *unfit*.

**Body**

if ( $z_1 = z_2 = z_3$ ) then

if total number of positive classes  $\geq 2$

then  $q_i \leftarrow \text{fit}$

else  $q_i \leftarrow \text{unfit}$ ;

else if ( $z_1 = z_2$ ) then

if class. $c_1$  = class. $c_2$  then  $q_i \leftarrow \text{class}.c_1$

else  $q_i \leftarrow \text{class}.c_3$ ;

else  $q_i \leftarrow \text{class}.c_1$ ;

**End-Body**

**End-Pseudocode**

#### 6.2.4 Stage 4: Ranking candidate researchers

Having closed the *purpose-oriented classifier* cycle, in *Stage 4* a rank of the most similar candidate CVs classified as *fit for the purpose* is generated to assist decision makers in the assessment processes such as recruitment, promotion and funding. This rank of candidate researchers is produced considering the similarity between the researchers of example CVs (i.e., positive cases). These resulting rank, delivered to decision makers, can be reused and adapted in a new cycle of this CBR implementation.

### 6.3 CONCLUDING REMARKS

Chapter 6 introduced the *purpose-oriented method* to answer the main research question of this thesis, “*How to assess researcher quality for collaborative purposes?*”. The proposed method taken into account that the researchers’ career trajectories are experiences mapped into CVs. Thus, the similarity between a new candidate researcher and a successful researcher is assessed through the attributes of their CVs.

The approach initiates by defining fundamental inputs for the *purpose-oriented classifier*, which are: The Leiden Manifesto principles (HICKS et al., 2015); Knowledge Engineering methods, particularly, the CBR methodology and ML methods. The first is a mature methodology with a vast availability of methods, which may be combined in each step of the CBR cycle. The second is responsible by engineering computational solutions supporting the approach; In addition, the Brazilian *Lattes* database, which was chosen as data source.

The core of the proposed method is the *purpose-oriented classifier*, which is an implementation of similarity heuristics. This proposed classifier is tackled in four stages as illustrated in Figure 28. In *Stage 1*, the quality requirements for the assessment process is established by decision makers through example CVs; In *Stage 2*, the purpose of the assessment is learned through weights; In *Stage 3*, candidate researchers are classified as *fit* or *unfit* for the purpose of the assessment; and In *Stage 4*, candidate researchers are finally ranked.

By executing these four stages the *purpose-oriented method* automatically assesses the researcher quality in selection processes, such as recruitment, promotion, or grant awarding decisions. In chapter 7, this proposed method will be demonstrated in detail, in two collaborative experimental scenarios.

## 7 USEFULNESS OF THE METHOD

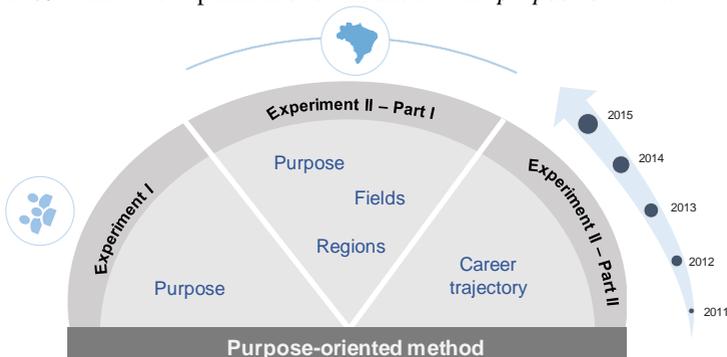
### 7.1 INTRODUCTION OF THE CHAPTER

The goal of Chapter 7 is to demonstrate the usefulness of the *purpose-oriented method* through different experimental scenarios of science and technology. To this end, two experiments were created, which will be briefly described in this introduction, and detailed in the next sections.

The first experiment introduces the *purpose-oriented method*, focusing on the purpose of the assessment. This experiment is based on the hypothesis that *a purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method*. Then, to test the experiment, a scenario is created considering a small sample of 80 CVs, from the *Lattes* database, which are artificially labeled as *fit* or *unfit* for collaborative purposes.

The *second experiment* effectively demonstrates the *purpose-oriented method* in a real example using data from a Brazilian research group. This experiment is divided in two parts: The first one, taking into account the results of Experiment I, presenting the *purpose-oriented method* in detail, focusing on the current candidate researcher production. In the second part, the treatment of career trajectories is incorporated into the proposed method based on the assumption that the *purpose-oriented method* could produce results better aligned with goals of the assessment. Thus, the researcher production is examined in a target interval of candidate researchers' career trajectory. Figure 39 illustrates these two experiments performed in Chapter 7.

Figure 39 – The two experiments to demonstrate the *purpose-oriented method*



Source: The author, 2017.

## 7.2 EXPERIMENT I

### 7.2.1 Introduction

The first experiment demonstrated in this chapter, introduces *purpose-oriented methods* to assess researcher quality. This experiment considers the hypothesis:

*“H2.1: A purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method”.*

The intention is to demonstrate that, on average, a classifier trained with data using quality references tailored to the peculiarities of their context (i.e., a specific purpose) is more accurate than a classifier that does not consider the context of its target selection process (i.e., a purpose-independent). The approach adopted relies on the second principle of the Leiden Manifesto (HICKS et al, 2015), which suggests paying attention to the context and goals of the institutions, groups or researchers, when assessing researcher quality. Taking this principle into account, this approach integrates purpose with information about the selection process that needs to assess researcher quality.

In order to demonstrate the hypothesis aforementioned, an application scenario is created with two different selection processes. The first selection process considers fit, researchers who are successful in collaborative productions, and thus are expected to succeed in collaborative research. The second selection process considers fit, researchers who are successful in solo productions, and thus are expected to succeed in solo research.

### 7.2.2 Data

This experiment uses a sample of 115 researcher CVs, selected randomly, in 2015, from the Brazilian *Lattes* database (lattes.cnpq.br), which was described on Section 2.4.7.1., and whose attributes and data representation were described in Chapter 5. Specifically, all attributes presented in the Frames 26 were used in the Experiment I.

After applying a data pre-processing, these 115 researcher CVs resulted in 84 CVs. After that, two datasets were created to represent each different target selection process. The first dataset, Collab, has 13 researcher CVs labeled as fit because their accomplishments are aligned with success in collaborative endeavors, and 15 researcher CVs labeled

as unfit. The second, Solo, has 14 researcher CVs with successful solo accomplishments that are labeled as *fit*, and 15 researcher CVs labeled as *unfit*.

Then, data from these two datasets (i.e., Collab and Solo) were aggregated to create a third one that does not distinguish a specific selection process. This third dataset has 57 researcher CVs without a defined purpose (i.e., purpose-independent).

The researcher CVs were labeled as *fit* or *unfit* by following consistent rules favoring *Collab* and *Solo* datasets, respectively. There is no intersection in these datasets. All *unfit* researchers were labeled as such, within the rules for each selection process. These three datasets were represented in a *feature vector representation*, as presented in Chapter 5, Section 5.5, Table 5.

Table 5 – The case base of CVs represented by a feature-value pairs data structure

Researchers	Article (a <sub>1</sub> )	Book (a <sub>2</sub> )	Patent (a <sub>3</sub> )	...	Grant (a <sub>n</sub> )	Class
Researcher 1 (r <sub>1</sub> )	10	3	1	...	3	fit
Researcher 2 (r <sub>2</sub> )	5	3	0	...	1	fit
Researcher 3 (r <sub>3</sub> )	40	9	2	...	10	unfit
...	...	...	...	...	...	unfit
Researcher m (r <sub>m</sub> )	15	5	1	...	2	fit

Source: The author, 2017.

The *Purpose-Independent dataset* includes all 27 fit researchers from Collab and Solo, and 30 added unfit researchers. This dataset does not distinguish any specific selection process. The three datasets are described in Table 9.

Table 9 – Number of *fit* and *unfit* researchers in each dataset

Dataset	<i>Fit</i>	<i>Unfit</i>	Total
Collab	13	15	28
Solo	14	15	29
Purpose-Independent	27	30	57

Source: Duarte, Weber and Pacheco (2016b)

### 7.2.3 Methodology

The purpose-oriented classifier, described on of the Section 6.2, was applied to demonstrates the hypothesis “H2.1: A purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method”. Thus, each dataset was firstly used as training data to learn weights to create three different classifiers. These classifiers are named, respectively, purpose-oriented collaboration classifier (POC1), purpose-oriented solo classifier (POC2), and purpose-independent classifier (PIC).

ReliefF (KNONENKO; ROBNIK-SIKONJA;POMPE, 1996) from Weka (<http://www.cs.waikato.ac.nz/ml/weka>) was applied to learn weights. In this approach, weights are used to represent the attribute relevance for a specific purpose in this CBR classification. Once each set of weights populates the classifier, it becomes a specific classifier.

The method adopted to validate the result was Leave-One-Out Cross-Validation (LOOCV) (e.g., GU; AAMODT, 2006; KOHAVI, 1995; REFAELZADEH; TANG; LIU, 2009), which represents the accuracy of the method, and is defined as the percentage of instances correctly classified by the algorithm. The set of CVs classified as fit or unfit is the ground truth, which is used as reference of accuracy.

### 7.2.4 Purpose-oriented or purpose-independent assessment?

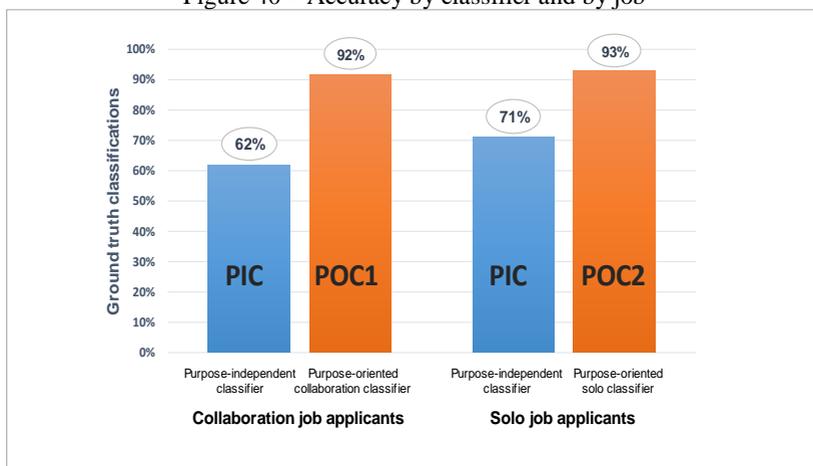
The *first purpose-oriented collaboration classifier* (POC1) is the classifier oriented to assess quality of applicants for the purpose of a job that seeks a collaborator, which we created hypothetically.

The *second purpose-oriented solo classifier* (POC2) is the classifier oriented to assess quality of applicants for the purpose of a job that seeks someone for solo work, which we created hypothetically.

The *purpose independent classifier* (PIC) is that considers universal standards to assess quality without contextual aspects of a specific purpose (e.g., a job opening).

The accuracy between the two *purpose-oriented* (POC1, POC2) is compared against the *purpose-independent classifier* (PIC), and the results are shown in Figure 40.

Figure 40 – Accuracy by classifier and by job



Source: Duarte, Weber and Pacheco (2016b)

The *purpose independent classifier* correctly classified 62% of the applicants for the collaboration job, and 71% for the solo job. The *purpose-oriented classifiers* correctly classified 92% and 93%, respectively. These results support the hypothesis that *purpose-oriented classifiers are more accurate than purpose independent*. These levels of accuracy of the *purpose-independent classifier* would falsely consider unfit five and four applicants, respectively, that are actually fit for the collaboration and solo jobs. The performance of the *purpose-oriented methods* would have falsely labeled only one applicant in each of the jobs.

Table 10 – Similarity score between candidate CVs to collaborative purposes

Candidate CVs (New cases)	Best similar CV		Second best similar CV		Third best similar CV	
	Best similar CV	Sim Score	Second best similar CV	Sim Score	Third best similar CV	Sim Score
Holmes	Tom	0,9357	Vilma	0,9289	Penelope	0,9281
Vilma	Tom	0,9297	George	0,9289	Beth	0,9263
Monica	Holmes	0,8888	Elroy	0,8639	Rosie	0,8591
George	Elroy	0,7929	Tom	0,7841	Rosie	0,7803
Marvin	Olivia	0,7438	George	0,6562	Fred	0,6411

Source: The author, 2017.

Table 10 illustrates the rank of the three first similarity scores considering applicants for the collaboration job, obtained after the classification process.

The table lists five candidate CVs (i.e., new cases) and their respective most similar example CVs (i.e., cases), which are fit researchers who are successful in collaborative purposes.

### **7.2.5 Concluding remarks**

In this experiment, the purpose-oriented method to assess researcher quality was introduced, described, and validated. The approach is aligned with the OECD (2008), which based on the notion that quality is fitness for purpose (JURAN; GODFREY, 1999), recommends incorporating purpose when constructing research metrics. It also implements the second principle of the Leiden Manifesto (HICKS et al, 2015), which argues that performance should consider contextual aspects.

The proposed approach considered contextual aspects through examples of researcher CVs, which were artificially labeled by humans as fit and unfit, following consistent rules. After that, the purpose was learned through weights applied to these contextual aspects. This experiment demonstrated the viability of the proposed method and allowed me to continue this investigation exploring other situations, which will be described in the next sections.

## **7.3 EXPERIMENT II**

### **7.3.1 Introduction**

The second experiment simulates a recruitment process of candidate researchers to work in collaboration with members of a research group. In general, this process involves the selection of researchers from different knowledge areas, institutions and cultures to work in common goal with other researchers. In order to simulate such scenario, the first challenge was to find a real example, which could be used as reference in this experiment. Thus, I investigated the *Microcephaly Epidemic*

*Research Group* (MERG)<sup>6</sup>, and found it to be a suitable Brazilian case undoubtedly referenced as being high-quality.

The *Microcephaly Epidemic Research Group* (MERG)<sup>7</sup> is established at the Aggeu Magalhães Research Center (CPqAM), a facility of the Oswaldo Cruz Foundation (Fiocruz), in Recife, Pernambuco. MERG was created in 2015 during a public-health emergency of microcephaly in the state of Pernambuco. Currently it is composed of 21 PhD researchers from several institutions from Brazil, the UK, and the USA, who are specialized in the fields of health and biology. This renowned research group was previously described in Section 2.3.6.3.

In this experiment, I assume that these 21 PhD researchers are unquestionably successful collaborators, who were selected to work together in a high-quality research group from Fiocruz (henceforth Fiocruz MERG). Consequently, their CVs are real data that compose a set of successful example CVs, that is, they are absolutely fit for the collaborative purpose of the Fiocruz MERG. Thus, in this experiment, the purpose-oriented method is applied using real data to predict whether a candidate researcher is fit or unfit for the purpose of the assessment.

This experiment is aligned to the Leiden Manifesto (HICKS et al, 2015), and concerns at least four of its principles. The second principle suggests aligning the metrics with the mission and purposes of institutions; the third principles emphasizes the need to acknowledge local instead of universal research; the sixth principle recommends taking into account the specificities of fields and publication practices; and the seventh principle suggests that individual researchers should be assessed based on a qualitative judgement of their portfolio.

Experiment II is organized in two parts, in the first part the purpose-oriented method is effectively demonstrated, focusing on the first three principles afore mentioned. The second part focuses specifically on the seventh principle, incorporating the treatment of career trajectory of researchers into the proposed method.

### 7.3.2 Data

In order to set this application scenario, data were extracted from the Brazilian *Lattes* database (lattes.cnpq.br). The process of data extraction, selection the attributes, and how these data are represented in

---

<sup>6</sup> The Microcephaly Epidemic Research Group (MERG):

<http://www.cpqam.fiocruz.br/merg/>

<sup>7</sup> CNPq Research Group: <http://dgp.cnpq.br/dgp/espelhogrupo/2723404431935999>

the *proposed purpose-oriented method*, are described in Chapter 5. Specifically, all attributes presented in the Frames 24, 25 and 26 were used in the Experiment II.

Thus, by considering the data source extracted in July 2016, CV data from each member of the Fiocruz MERG were selected. The resulting dataset includes 20 researcher CVs, the majority of whom, are from two CNPq knowledge areas (i.e., health sciences, and biological sciences), with accomplishments registered from 1989 to 2015. This dataset is called the *set of example CVs*.

After that, the *set of candidate researchers* was created, including researchers from all five regions of Brazil (i.e., South, Southeast, Midwest, Northeast, and North), and the same CNPq knowledge areas of the Fiocruz MERG. This dataset contains 15,266 researcher CVs from 1950 to 2015, which is called the *set of candidate CVs*.

Table 11 shows the distribution of researchers from the *set of example CVs* and the *set of candidate CVs* in the CNPq knowledge areas. This distribution considers that a researcher can be affiliated to more than one area. As illustrated, the majority of researchers are affiliated to health sciences.

Table 11 – The set of example CVs and set of candidate CVs by CNPq knowledge areas

<b>Knowledge area</b>	<b>Set of example CVs</b>	<b>Set of candidate CVs</b>
Health sciences	18	12428
Biological sciences	4	3296
Other	1	8

Source: Data extraction process for this thesis. The author, 2017.

Table 12 details Table 11, illustrating the distribution of researchers from the *set of example CVs* and the *set of candidate CVs* in CNPq knowledge sub-areas. For each knowledge area, the table shows the specialties of the member of Fiocruz group, evidencing the interdisciplinarity of the group. Collective health and Medicine are the sub-areas with the largest number of researchers, and hence with the largest number of candidate researchers

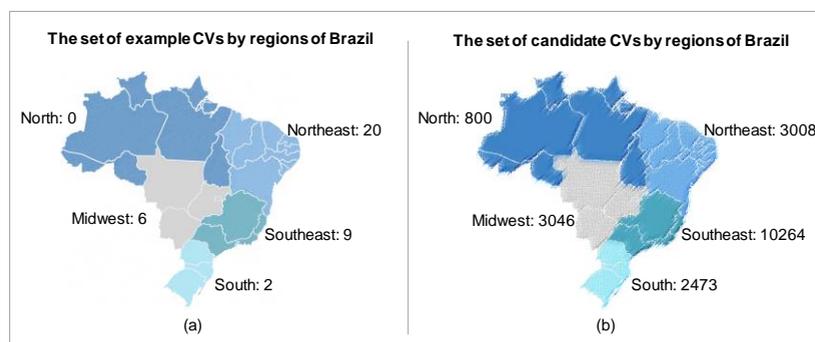
Table 12 – The set of example CVs and set of candidate CVs by CNPq knowledge sub-areas

Knowledge area	Knowledge sub-area	Set of example CVs	Set of candidate CVs
Health sciences	Collective Health	13	6736
	Medicine	11	5439
	Nutrition	3	727
	Phonoaudiology	1	593
Biological sciences	Parasitology	2	1468
	Immunology	2	779
	Microbiology	3	630
	Pharmacology	1	562
Other	Hospital administration	1	8

Source: Data extraction process for this thesis. The author, 2017.

Figure 41 illustrates the geographic distribution of researchers from the *set of example CVs* (Figure 41-a) and the *set of candidate CVs* (Figure 41-b), in the five regions of Brazil.

Figure 41 – The set of example CVs and set of candidate CVs by regions of Brazil



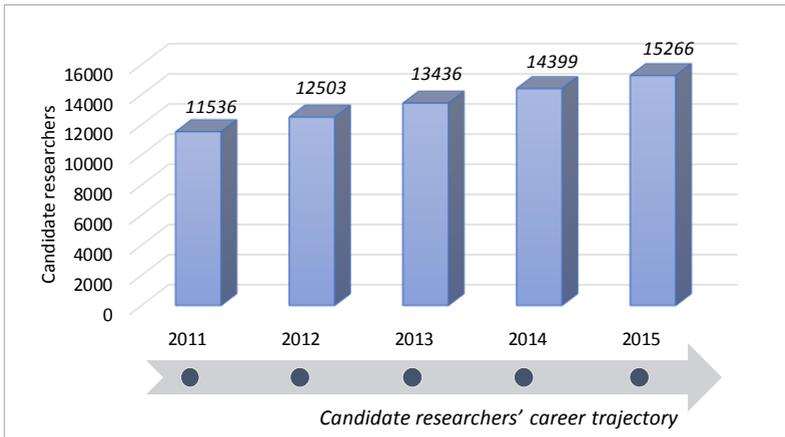
Source: Data extraction process for this thesis. The author, 2017.

In this geographic distribution, I considered the address of the institution in which a researcher is affiliated, taking into account that a researcher can be affiliated to more than one institution. For example,

Figure 41-a shows that the 20 members of the Fiocruz MERG, apart from the Northeast region, are affiliated to institutions from three other regions of Brazil.

At last, Figure 42 illustrates the candidate researchers' accomplishments along their career trajectory, distributed in five sets of candidate CVs, in a target interval of time of five years, from 2011 to 2015.

Figure 42– The set of candidate CVs from the 2011 to 2015



Source: Data extraction process for this thesis. The author, 2017.

The set of example CVs was previously labeled as the class *positive* ( $P$ ), and defined as the positive set  $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_n)$ . The set of candidate CVs was previously labeled as the *class unlabeled* ( $U$ ), and defined as the unlabeled set  $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ . The sets ( $P$ ) and ( $U$ ) have the same structure, and they are organized as a *feature-value pairs* data structure, where each row describes a researcher ( $r_n$ )  $\in R$ , with  $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_n)$ , and each column represents one of their attributes ( $a_m$ )  $\in A$ , with  $\mathbf{A} = (\mathbf{a}_{11}, \dots, \mathbf{a}_{mk})$ . The value of each cell of the table represents the absolute value of productivity of a researcher in each type of attribute. The list of attributes was described in Section 5.3, and the feature-value pairs data structure is represented as in Table 6, Section 6.2.2.1.

Figure 43– The data representation of a candidate researchers’ career trajectory

Researcher	Year <sub>1</sub>	Year <sub>2</sub>	...	Year <sub>N</sub>
	$a_{ij1}$	$a_{ij2}$	...	$a_{ij5}$
$r_1$	1	2	...	5
$r_2$	5	4	...	1
...	...	...	...	...
$r_n$	0	0	...	3

The production in a year

Source: Duarte, Weber and Pacheco (2016c)

The career trajectory of candidate researchers is represented by the set  $T = (t_1, \dots, t_N)$ , where each  $t_i$  is composed of the *positive set* ( $P$ ) and the *unlabeled set* ( $U$ ) in a given year, and  $N$  represents the length of a candidate researcher’s career trajectory. Figure 43 illustrates the data representation of a candidate researchers’ career trajectory

### 7.3.3 Methodology

Experiment II is conducted in order to effectively present the *purpose-oriented method* in detail. The intention is to apply the four stages of the proposed method, and to demonstrate the assumption:

“A2.1: *Incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment*”.

To this end, the experiment is divided in two parts: The first part applies the *purpose-oriented method*, but does not consider the career trajectory of researchers. In contrast, the second part, applies the proposed method considering a target interval of candidate researchers’ career trajectory. In this case, the treatment of career trajectories is incorporated into the proposed method. At the end, the results of the two parts of the experiment are analyzed and compared. These two parts of experiment II will be outlined next.

***Experiment II – Part I***

1. Apply the *purpose-oriented method*, which was described in Chapter 6, using the *set of example CVs*, and the *set of candidate CVs*, as input to perform the stages:
  - i. Stage 1: Describing the problem.
  - ii. Stage 2: Learning weights to represent purpose.
  - iii. Stage 3: Classifying candidate researchers as fit or unfit for the purpose.
  - iv. Stage 4: Ranking researchers.

***Experiment II – Part II***

1. Apply stages 1 to 3 of the *purpose-oriented method*, for each year of the target interval, using the set of example CVs, and the sets of candidate CVs.
  - i. Stage 1: Describing the problem
  - ii. Stage 2: Learning weights to represent purpose
  - iii. Stage 3: Classifying candidate researchers as *fit* or *unfit* for the purpose of the assessment
2. Rank candidate researchers considering the target interval of career trajectory
3. Analyze whether the treatment of career trajectories incorporated into the *purpose-oriented method* will produce results better aligned with the goals of the assessment.

**7.3.4 Experiment II – Part I**

In this subsection, following the four stages of the propose method, a selection process of candidate researchers to work in collaboration with members of a research group is simulated.

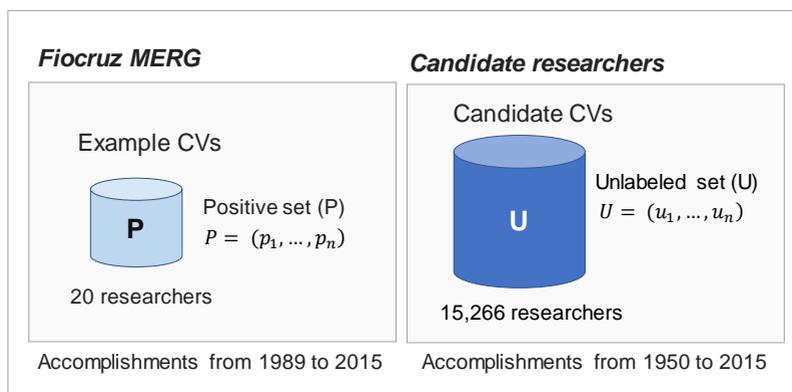
***Stage 1: Describing the problem***

Given the problem previously stated on Section 7.3.1, a resulting setup is summarized for this first part of this experiment, in the following items:

- i. The problem: How to assess the quality of candidate researchers to work in collaboration with members of the Fiocruz MERG.
- ii. Experimental scenario: The Fiocruz MERG
- iii. *Source of principles for assessing researcher quality*: The Leiden Manifesto principles (HICKS et al, 2015):

- *Principle 2*: Measure performance against the research missions of the institution, group or researcher;
  - *Principle 3*: Protect excellence in locally relevant research.
  - *Principle 6*: Account for variation by field in publication practices.
- iv. *Knowledge Engineering: The CBR methodology and ML methods*
- v. *Data source: The Brazilian Lattes database*
- vi. *Resulting datasets: As illustrated in Figure 44.*

Figure 44 – The resulting datasets of Experiment II – Part I



Source: Data extraction process for this thesis. The author, 2017.

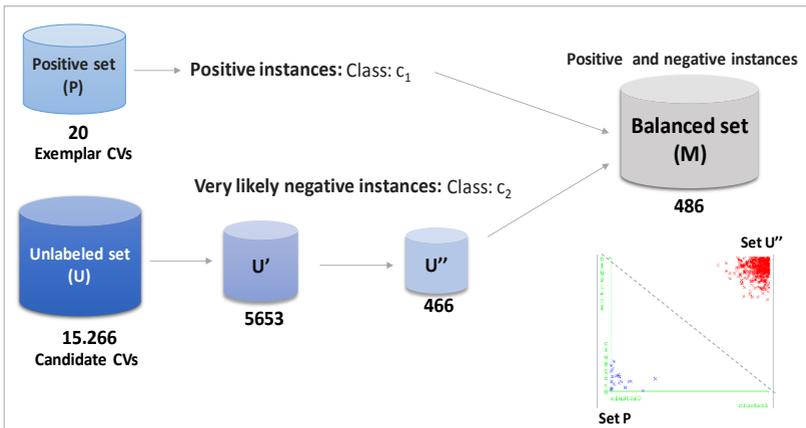
### ***Stage 2: Learning weights to represent purpose.***

Stage 2 is illustrated by demonstrating step by step how the proposed method represents the purpose of the assessment through learning weights. Since the quality requirements of this application scenario contain only successful CVs, the problem faced here is the lack of negative instances to learn weights. Thus, the Spy technique in S-EM (LIU et al., 2002) is used to create negative instances, as described in Section 6.2.2.1, in the following subsections.

1. *Initializing the balanced set of both positive and negative instances*

The positive set (P) and likely negative instances from the unlabeled set (U) were integrated in a balanced set (M). The set of likely negative instances is a sample of instances from the set (U) completely opposite to the instances from the set (P). The process to generate the set (M) first calculated the sum of each attribute of the set (P) (i.e.,  $\text{sum}(a_i) = \sum a_i, i=1\dots n$ ). After that, each attribute  $a_i$  was analyzed, and a threshold ( $t$ ) = 10 was used to decide whether an attribute had higher productivity (H) or lower productivity (L). Then, a filter was applied to find instances in the set (U) using the subset of attributes with lower productivity (L). This process resulted in the set  $U' \leftarrow (I_1 = 0 \text{ and } I_2 = 0 \text{ and } I_k = 0)$  with 5,653 instances. At last, using this result to obtain a more balanced set, instances whose productivity was equal to 1, 10 and 15 accomplishments, were selected. This last step resulted in the set ( $U''$ ) with 466 likely negative instances. Finally, a balanced set (M), of both positive and negative instances, was composed with instances from the set P and the set ( $U''$ ). Figure 45 illustrates the process of initialization of the balanced set (M).

Figure 45 – The process of initialization of the balanced set (M)



Source: The author, 2017.

## 2. Applying the Spy technique in S-EM (LIU et al., 2002)

Having initialized the balanced set (M), the Spy technique in S-EM (LIU et al., 2002) was applied, as described in Section 6.2.2.1. At the end, the resulting probability of a CV belong to class  $c_1$  is sorted from the highest to the lowest, as illustrated in Figure 46.

Figure 46 – The alignment between the example CVs

#	pCluster_0_0	%	
1	9	100%	genuine successful CV (i.e. positive)
3	9	100%	
4	9	100%	
8	9	100%	
9	9	100%	
10	9	100%	
14	9	100%	
17	9	100%	
16	9	100%	
5	8	89%	uncertain CV
12	7	78%	
13	7	78%	
11	7	78%	
15	7	78%	
2	6	67%	Outliers
7	6	67%	
18	6	67%	
20	6	67%	
19	5	56%	
6	2	22%	
Tot	9		

Source: The author, 2017.

Figure 46, shows that the researchers of the set of example CVs are aligned in three groups: genuine successful CVs, uncertain CVs, and outliers. For example, a group of nine researcher CVs were identified as genuine successful CV. Another group, also with nine researcher CVs are a little distant, however, the in third group are two researcher CVs identified as outliers, which for some unknown reason have a behavior much different from the two other groups.

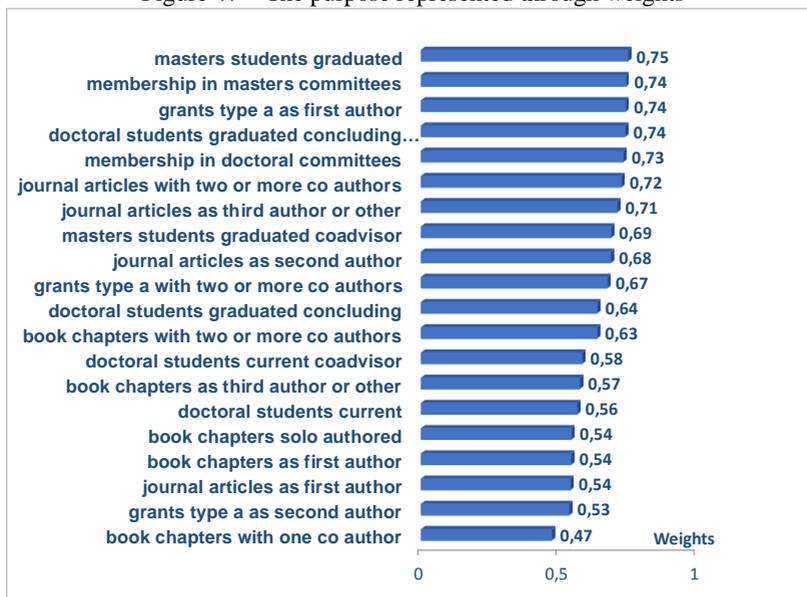
## 3. Applying feature weighting to represent purpose

After categorizing the exemplar CVs in genuine successful CVs, uncertain successful CVs, and outliers, the next step was to characterize the context of the set of exemplar CVs, that is, to represent the purpose of the assessment through the relative relevance of attributes from the set of example CVs. This was performed by applying feature weighting methods, whose results are illustrated in Figure 47.

In order to apply feature weighting, a set  $M' = (m'_1, \dots, m'_n)$  was created, which include two sets: the set  $P' = (p'_1, \dots, p'_n)$  that is composed of the positive set (P) without those instances that are outliers, and the set of likely negative instances  $U' = (u'_1, \dots, u'_n)$ . Instances from the set ( $P'$ ) were labeled as positive and instances from the set ( $U'$ ) where labeled as negative. The final set ( $M'$ ) has 486 instances, distributed in 18 instances labeled as positive and 468 instances labeled as negative.

In sum, the previously mentioned problem of the lack of negative instances to learn weights was solved, and after that, three *feature weighting algorithms* were applied: Correlation Based Feature Selection (CFS) (HALL, 1999), Information Gain (IG) (QUINLAN, 1992), and ReliefF (KIRA; RENDELL, 1992; KNONENKO; ROBNIK-SIKONJA; POMPE, 1996), by using Weka tools (<http://www.cs.waikato.ac.nz/ml/weka/>). At the end, two of these algorithms outputted similar results, the CFS and the IG.

Figure 47 – The purpose represented through weights



Source: The author, 2017.

Figure 47 shows the 20 most relevant attributes that represent the context of the group, and their respective weights, ranked by the CFS

algorithm. The picture reflects at least two trends present in the CVs of member of the Fiocruz MERG. The first trend relates to activities in education, for example, mentoring master and doctor students. Other attributes, such as journal article and book chapters relate to scientific activities. That means the purpose of this assessment process is to select candidate researchers that have accomplishments in education and publications of scientific research.

### **Stage 3: Classifying candidate researchers as *fit* or *unfit* for the purpose.**

In stage 3, each step of the CBR implementation for the purpose-oriented classifier is demonstrated, as described below.

#### *1. Initializing the datasets of cases, new cases, and weights*

The set of cases ( $C$ ) is defined as  $C = (c_1, \dots, c_m)$ , and composed of 486 instances, which are 20 positive instances and 466 negative instances. The set of new cases ( $Q$ ) is defined as  $Q = (q_1, \dots, q_n)$ , and composed of 14800 unlabeled instances, 15266 instances from the set of candidate CVs minus the sample of 466 likely negative instances. The set of weights is defined by the set  $W = (w_1, \dots, w_m)$ , where each  $w_i$  is a value between  $[0,1]$ . This set was populated with the weights learned using the CFS algorithm (HALL, 1999).

#### *2. Calculating the score of similarity between the new cases ( $Q$ ) and cases ( $C$ )*

The global similarity **Global Sim** ( $q_i, c_j$ ) is calculated comparing each new case ( $q_i$ ) to all cases ( $c_1, \dots, c_m$ ), resulting in a rank of the nearest neighbours  $NN(q_i, c_{j,k}) = (z_1, \dots, z_n)$ .

#### *3. Reducing negative instances to obtain a more balanced set of cases*

Firstly, the set of nearest neighbours to each new case ( $q_i$ ) was ordered in a neighbour ranking, from the highest to lowest similarity score. After that, the strategy to reduce the redundant negative cases was applied. The task of reducing negative cases could have been executed in step1, when initializing the unbalanced set of cases ( $C$ ), which contains

20 positive cases and 466 likely negative cases. However, I opted to keep a larger diversity of negative instances when applying the similarity measure.

Then, in this step, a more balanced set is needed to identify the most suitable solution to the next step. In the set of cases (C), 20 cases are authentic positive instances, which were previously analyzed, and only two outliers were found, and for this reason they cannot be reduced. The 466 remaining cases are likely negatives, and hence, the negative instances should be reduced.

#### 4. *Classifying a new case as fit or unfit for the purpose of the assessment*

Having the final nearest neighbours ranked, the proposed rule presented in Pseudocode-1 was applied considering the first three *nearest neighbours*. Table 13 illustrates this resulting classification in a sample of 10 new cases classified as *fit* or *unfit* for the purpose of the assessment.

Table 13 – The resulting classification of the Experiment II – Part I

Candidate CVs		Example CVs								
New case		1 <sup>st</sup> most similar case			2 <sup>nd</sup> most similar case			3 <sup>rd</sup> most similar case		
#	Class	#	Class	Sim	#	Class	Sim	#	Class	Sim
10	unfit	355	N	0,9997	457	N	0,9996	238	N	0,9995
22	fit	78	P	0,9851	371	N	0,9836	343	N	0,9831
23	fit	348	P	0,9169	149	P	0,9029	109	P	0,8930
34	unfit	412	N	0,9937	190	N	0,9936	114	P	0,9934
37	fit	114	P	0,9994	450	N	0,9992	336	N	0,9991
41	fit	310	P	0,9771	207	P	0,9761	81	P	0,9712
51	unfit	465	N	0,9852	329	N	0,9846	310	P	0,9843
54	fit	78	P	0,9841	207	P	0,9833	320	N	0,9828
72	fit	308	P	0,9794	68	P	0,9787	233	P	0,9779
236	fit	114	P	0,9953	190	N	0,9953	139	N	0,9952

Source: The author, 2017.

It is interesting to observe that the proposed rule used to execute the classification considers the class, the similarity score, and the sort order of the three cases. For example, analyzing the candidate CVs 22, 37 and 236, could lead to the understanding that such candidates should be classified as *unfit*, because there are two negative cases. However, they were classified as *fit* for the purpose, because the 1<sup>st</sup> most similar case is positive.

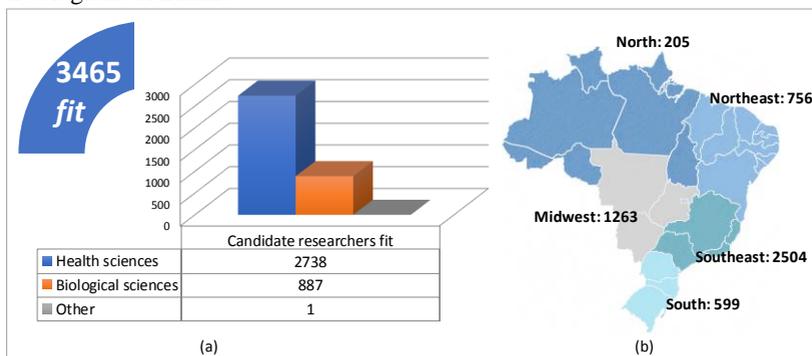
## 5. Presenting the classification results

The classification process resulted in 3465 candidate CVs classified as *fit* and 11335 candidate CVs classified as *unfit* for the purpose of the assessment. Then, taking into account these results, I started by contextualizing *the candidate CVs classified as fit*, by fields and regions. as illustrated in Figure 48.

As shown in Figure 48-a, health sciences was the CNPq knowledge area with the largest numbers of candidate researchers classified as *fit*. It is also important to highlight that at least one candidate researcher classified as *fit* belongs to the CNPq knowledge area “other”, in the hospital administration subarea.

Figure 48-b presents the geographical distribution of candidate researchers classified as *fit*, in the five regions of Brazil. The Southeast, had the largest number of candidate researchers selected as *fit*, followed by the Midwest, and Northeast.

Figure 48 – The fit candidates distributed in the CNPq knowledge areas and the five regions of Brazil.



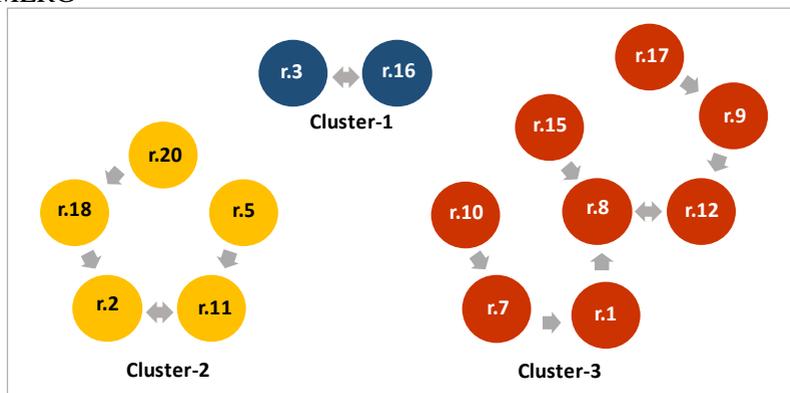
Source: The author, 2017.

After the contextualization, considering that the CVs of the members of Fiocruz MERG also were among the set of candidate CVs, I analyzed the *1<sup>st</sup> most similar case*, in order to understand the similarity between them. As a consequence, an impressive result emerged, which showed three distinct clusters of similarity, as illustrated in Figure 49.

These clusters reveal three different profiles of researchers inside the same group. *Cluster-1 (blue)* is the smallest of them, and shows that researchers r.3 and r.16 are similar to each other. *Cluster-2 (yellow)* has five researchers, and the main core are r.2 and r.11, which are most similar

to each other. Cluster-3 (red) is the biggest, with eight similar researchers. In this cluster, the researchers r.8 and r.12 are most similar to each other, and because three other researchers are directly linked to r.8, it is the core of this cluster-2.

Figure 49 – Clusters of similarity between the members of Fiocruz MERG

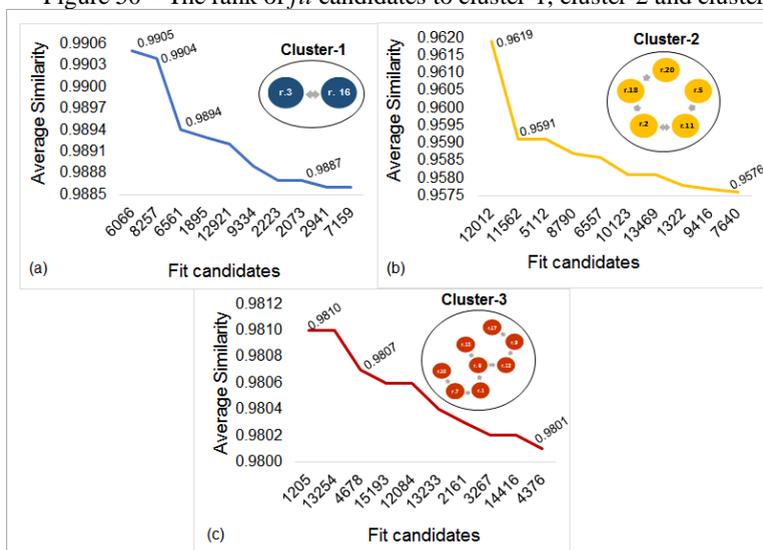


Source: The author, 2017.

Thus, the *purpose-oriented classifier* provides several veins of analysis, from these clusters of similarity, to decision makers. In the next step, the rank of candidate researchers fit for the purpose of the assessment is generated considering each cluster. Other analysis will be performed in the next section, in which I trace the evolution of the similarity between the members of the Fiocruz MERG along the last five years of their career trajectory. In future studies, I intend to investigate the interdisciplinary coproduction between members of a research group in more detail.

#### Stage 4: Ranking candidate researchers

The process of ranking candidate researchers for the purpose of the assessment considered the most similar *fit candidates* to each cluster. It first calculated the average similarity score of positive cases to each *fit candidate*. After that, it ranked the *fit candidates* to each cluster, by calculating again the average similarity score of each *fit candidate* to the members of the cluster. Figure 50 illustrates the resulting rank of *fit candidates* to each cluster.

Figure 50 – The rank of *fit* candidates to cluster-1, cluster-2 and cluster-3

Source: The author, 2017.

The chart of Figure 50-a shows the average similarity of the ten most similar *fit candidates* to members of cluster-1. For example, the *fit candidate* number 6066 is the most similar to researchers r.3 and r.16. from the Fiocruz MERG, with average similarity of 0.9905. Likewise, the charts of Figures 50-b and 50-c present respectively, the rank of the ten most similar *fit candidates* to members of cluster-2 and cluster-3. In this case, the *fit candidate* number 12012 is the most similar to cluster-2, with average similarity of 0.9619, and the *fit candidate* number 1205 is the most similar to cluster-3, with average similarity 0.9810.

Table 14 resumes the statistical results of the three clusters, presenting the number of *fit candidates* to respective cluster, and their maximum and minimum similarity scores. It is important to emphasize that the members of clusters were not included in the *fit candidates*.

Table 14 – The statistical results of cluster-1, cluster-2 and cluster-3 ranks

Cluster	Number of Fit candidates	Maximum similarity score	Minimum similarity score
Cluster-1	3463	0.9905	0.7903
Cluster-2	3460	0.9619	0.7992
Cluster-3	3457	0.9810	0.7939

Source: The author, 2017.

At the end of the process, a resulting rank is generated by calculating the average similarity of the fit candidates to the three clusters. Table 15 illustrates this resulting rank, showing the ten most similar fit candidates. The process generates a rank of 3450 fit candidates, where the maximum similarity score is 0.9705, and the minimum similarity score is 0.7974.

Table 15 – The resulting rank of *fit candidates* to the purpose of the assessment, in which the experiment does not considers career trajectory.

Rank	Fit candidate	Similarity Cluster-1	Similarity Cluster-2	Similarity Cluster-3	Average similarity
1	1205	0.9872	0.9433	0.9810	0.9705
2	13254	0.9865	0.9431	0.9810	0.9702
3	13834	0.9866	0.9440	0.9780	0.9695
4	1920	0.9858	0.9436	0.9791	0.9695
5	10959	0.9832	0.9469	0.9785	0.9695
6	9661	0.9853	0.9458	0.9772	0.9694
7	5568	0.9858	0.9431	0.9793	0.9694
8	11719	0.9849	0.9439	0.9789	0.9692
9	3526	0.9797	0.9481	0.9794	0.9691
10	76	0.9825	0.9464	0.9784	0.9691

Source: The author, 2017.

#### 7.3.4.1 Concluding Experiment II – Part I

In this first part of Experiment II, a recruitment process of candidate researchers was simulated by applying the proposed method step-by-step. In this simulation process, the problem of *lack of negative instances to learn weights* was faced, and the Spy technique in S-EM (LIU et al., 2002) was used to create negative instances. This problem was solved, and *feature weighting algorithms* was applied, in order to represent the purpose of the assessment through weights.

Characterizing the purpose of the assessment through learning methods was one of the most relevant results delivered by the *purpose-oriented method*. For instance, in this simulated process, the purpose identified was selecting researchers that had accomplishments in education and scientific publications to work in collaboration with member a research group.

After that, a CBR implementation was applied to classify candidate researchers as *fit* or *unfit* for the purpose of the assessment. Out of the 15,266 candidate researchers, 3,465 were classified as *fit* and 11,335 were classified as *unfit* for this simulated recruitment.

Then, I investigated how similar the members of the target group were to each other, taking into consideration that their CVs were also among the set of candidate researchers, and that they have been classified as *fit*. Hence, an impressive result emerged, revealing three distinct clusters of similarity, that is, three different profiles of researchers inside the same group. This was the second most relevant result of the *purpose-oriented method*, because this result demonstrated that the proposed method was able to identify the similarities within the group.

Finally, I ranked the fit candidates, firstly considering the most similar fit candidates to each cluster, by calculating the average similarity score of positive cases to each fit candidate. After that, I ranked the fit candidates to each cluster, by calculating again the average similarity score of each fit candidate to the members of the cluster. At the end of the process, a resulting rank was generated by calculating the average similarity of the fit candidates to the three clusters. The final result was a rank of 3,450 fit candidates.

Concluding, the goal to demonstrate the *purpose-oriented method* was achieved, and the proposed method was able to simulate a recruitment process of candidate researchers to work in collaboration with members a research group. Specifically, it was able to characterize the purpose of the assessment through weights; to classify candidate researchers as *fit* or *unfit* for the purpose of the assessment; and rank the *fit candidates* in a list, to assist decision makers in their assessment endeavors.

In the next section, the *purpose-oriented method* is demonstrated considering the career trajectory of researchers.

### 7.3.5 Experiment II – Part II

In this section, the purpose oriented method to assess researcher quality is conducted considering the career trajectory of researchers, as recommended in the seventh principle of the Leiden Manifesto (HICKS et al, 2015, p.430), “*Even when comparing large numbers of researchers, an approach that considers more information about an individual’s expertise, experience, activities and influence is best*”.

### 7.3.5.1 *Incorporating treatment of career trajectories into the purpose-oriented method*

The intention here is to demonstrate the assumption: “A2.1: Incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment”.

To this end, based on the application scenario created for the first part of this experiment (i.e., the Fiocruz MERG), five hypothetical scenarios for each of the last five years of the candidate researchers’ career trajectory, were created using the absolute values of productivity. It is also important to highlight that such candidate researchers were likely subject to the same conditions of scholarly research.

At the Twenty-Forth International Conference on Case-Based Reasoning (ICCB2016), in the Workshop on Reasoning about Time in CBR, another work published while developing this thesis, I investigated how to consider career trajectories to assess researcher quality (DUARTE; WEBER; PACHECO, 2016b). This work introduced an approach to preprocess data from CVs, based on the assumption that assessing researcher quality ultimately implies predicting future success. The study proposes strategies to compare researchers whose scholarly production is achieved under different conditions and career trajectories lengths, through different periods of time.

Thus, the same tasks performed in the first part of this experiment are applied in this second part to each set of candidate CVs of the target interval. At the end, the results of each year are consolidated in a unique rank of candidate researchers fit for the purpose of the assessment. This second part of the experiment will be detailed in the following subsections.

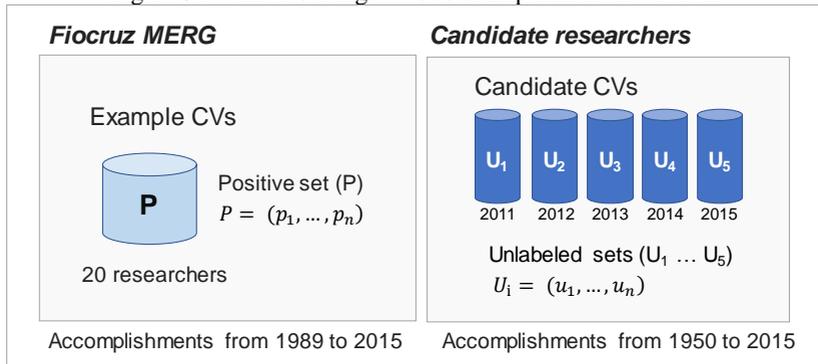
#### **Stage 1: Describing the problem**

Given the problem previously stated above, a resulting setup is summarized next:

- i. *Assumption:* as above mentioned
- ii. *Experimental scenario:* The Fiocruz MERG
- iii. *The target interval of career trajectory:* five years
- iv. *Source of principles for assessing researcher quality:* The Leiden Manifesto seventh principle (HICKS et al, 2015)
- v. *Methods of Knowledge Engineering:* The CBR methodology and ML methods

- vi. *Data source*: The Brazilian *Lattes* database
- vii. *Resulting datasets*: As illustrated in Figure 51.

Figure 51 – The resulting datasets of Experiment II – Part II



Source: Data extraction process for this thesis. The author, 2017.

## Stage 2: Learning weights to represent purpose.

In this second part of Experiment II, the weights previously learned in the first part are used to represent the purpose of the assessment, which is the hypothetical recruitment process of candidate researchers to work in collaboration with members of the Fiocruz MERG.

## Stage 3: Classifying candidate researchers as fit or unfit for the purpose.

The approach used to classify the set of candidate CVs is such as described in the first part of this experiment. These following steps are executed for each year of the target interval. Table 16 presents the resulting classification.

- i. Initialize the balanced set of both positive and negative instances
- ii. Calculate the score of similarity between the new cases (Q) and cases (C)
- iii. Reduce negative instances to obtain a more balanced set of cases
- iv. Classify a new case as *fit* or *unfit* for the purpose of the assessment

Table 16 – The resulting classification of the Experiment II – Part II

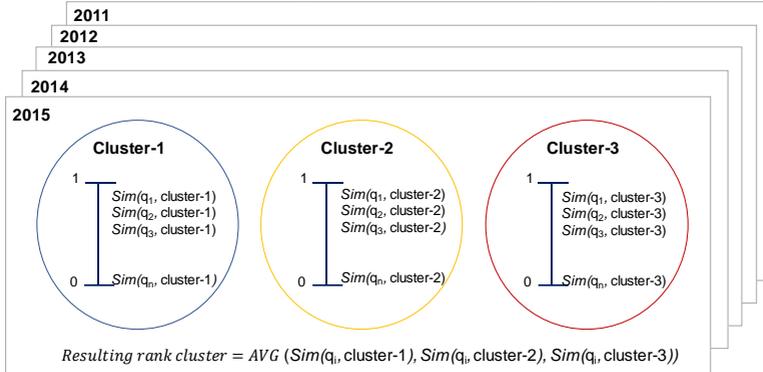
Candidate CVs	Target interval of career trajectory				
	2011	2012	2013	2014	2015
<i>Fit</i> candidates	2018	2301	2650	3072	3465
<i>Unfit</i> candidates	9062	9749	10321	10886	11335

Source: The author, 2017.

### Stage 4: Ranking candidate researchers

In order to rank fit candidates considering their career trajectory, two steps are performed: In the first, the same process described in Experiment II – Part I is applied to each year of the target interval.

Figure 52 – The resulting rank clusters for each year the target interval



Source: The author, 2017.

Then, a resulting rank cluster is generated consolidating the results of the three clusters in a unique rank, for each year of the target interval. This resulting rank cluster is calculated by the average similarity of each fit candidate to the three clusters. Figure 52 illustrates this step.

In the second step, the intention is to identify the research candidates that remained fit for the purpose of the assessment during the five consecutive years of the target interval. Thus, a career trajectory rank is generated by calculating the average similarity score from 2011 to 2015. This career trajectory rank includes 1918 candidates, where the highest similarity score is 0.9667 and the lowest is 0.7954. Table 17 illustrates the final career trajectory rank, considering the five first fit candidates. The rank of 100 most *fit candidates* for the purpose of the Experiment II – Part II, is presented in Appendix A – Frame 31.

Table 17 – The final career trajectory rank

Rank	Fit Candidate	Similarity score					Average similarity 2011-2015
		2011	2012	2013	2014	2015	
1	7550	0.9630	0.9658	0.9675	0.9686	0.9686	0.9667
2	9782	0.9626	0.9654	0.9670	0.9676	0.9686	0.9662
3	6004	0.9612	0.9649	0.9669	0.9672	0.9686	0.9658
4	3987	0.9623	0.9657	0.9669	0.9672	0.9666	0.9657
5	10736	0.9616	0.9646	0.9664	0.9677	0.9677	0.9656

Source: The author, 2017.

### 7.3.6 Analysis

In this section, the analysis of Experiment II is presented, focusing on the assumption that “*incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment*”. Thus, firstly the results of the two parts of Experiment II, are highlighted, as shown in Table 18.

Table 18 – The results of Experiment II

Experiment II		Candidate CVs	Unfit Candidate	%	Fit Candidate	%
Part-I		14800	11335	77%	3465	23%
Part-II	2011	11080	9062	82%	2018	18%
	2012	12047	9746	81%	2301	19%
	2013	12971	10321	80%	2650	20%
	2014	13958	10886	78%	3072	22%
	2015	14800	11335	77%	3465	23%
	2011-2015		14800	12882	87%	1918

Source: The author, 2017.

The analysis began by comparing the results of the first part of experiment, which resulted in 3,465 fit candidates (i.e., 23%), with the results of the second part, in which after incorporating career trajectories into the *purpose-oriented method*, only 1,918 candidates (i.e.,13%) were classified as fit for the purpose of the assessment. Hence, this first analysis suggests that the classification process became much stricter than in the first part of the experiment.

Thus, the reasons why the *1,547 fit candidates were not classified as fit* in the entire period of the assessment was investigated. As result, at least three main reasons were found: First some candidates were new researchers, and thus, they did not have enough accomplishments in that year. Second, some others were classified as *unfit* for the purpose in some of the years of the target period. Third, some were classified as fit for the purpose, but not consecutively in the entire assessment period. These results are shown in Table 19, and exemplified next.

Table 19 – Distribution of *unfit candidates* from 2011-2015

Candidate researchers	2011	2012	2013	2014	2015
New researchers	35	17	3	0	0
Classified as unfit	1463	1191	850	450	0
Classified as fit	49	339	694	1097	1547

Source: The author, 2017.

It is important to emphasize that this experiment is a simulation of a researcher quality assessment under certain hypothetical circumstances. In this simulated process, the requirement is to consider the candidate classified as *fit* in the entire period. However, the decision maker could have considered to create a deflator to be included in the ranking for candidates in both situations, *fit* or *unfit* in not consecutive years of the target assessment.

### 7.3.6.1 Candidate researchers that were new researchers

Table 20 shows that, for example, the candidates 317 and 3029 were not classified as fit until 2012; the candidates 1763 and 6534 were not classified as fit from 2011 to 2013; and the candidate 1913 was not classified as fit until 2013.

Table 20 – New researchers in some period of the target interval

Researchers	2011	2012	2013	2014	2015
317	-	x	x	x	x
1763	-	-	-	x	x
1913	-	-	x	x	x
3029	-	x	x	x	x
6534	-	-	-	x	x

Source: The author, 2017.

### 7.3.6.2 Candidate researchers classified as unfit in some period

Table 21 illustrates five candidates that were classified in some period of the assessment as *unfit for the purpose*. For example, candidate 152 was unfit for four consecutive years; and candidate 5447 was an unfit candidate in 2011 and 2014.

Table 21 – *Unfit candidates* in some period of the target interval

Researchers	2011	2012	2013	2014	2015
40	x	x			
152	x	x	x	x	
220	x	x	x		
5447	x			x	
14787		x	x		

Source: The author, 2017.

### 7.3.6.3 Candidate researchers classified as fit in some period

Table 22 presents a sample of five candidate researchers from that were classified as *fit for the purpose* but only in some of the five consecutive years of the target interval.

Table 22 – *Fit candidates* in some period of the target interval

New researchers	2011	2012	2013	2014	2015
44					x
64				x	x
990	x				x
1138			x		x
7350		x			x

Source: The author, 2017.

### 7.3.6.4 Candidate researchers classified as fit in the entire period

Special attention is given to the analysis of the 1,918 who were *fit candidates* in all of the five consecutive years of the target assessment. This analysis considers the 10 most similar *fit candidates*, who were ranked by the average similarity score regarding the entire period, as shown in Table 23.

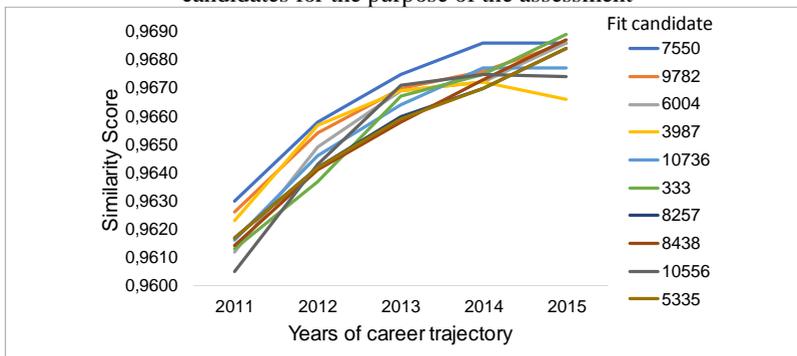
Table 23 – The 10 most similar *fit candidates* for the purpose of the assessment

Fit candidate	Similarity score by year					Similarity score rank 2011-2015
	2011	2012	2013	2014	2015	
7550	0.9630	0.9658	0.9675	0.9686	0.9686	0.9667
9782	0.9626	0.9654	0.9670	0.9676	0.9686	0.9662
6004	0.9612	0.9649	0.9669	0.9672	0.9686	0.9658
3987	0.9623	0.9657	0.9669	0.9672	0.9666	0.9657
10736	0.9616	0.9646	0.9664	0.9677	0.9677	0.9656
333	0.9613	0.9637	0.9667	0.9675	0.9689	0.9656
8257	0.9617	0.9642	0.9660	0.9670	0.9684	0.9655
8438	0.9614	0.9641	0.9658	0.9673	0.9687	0.9655
10556	0.9605	0.9643	0.9671	0.9675	0.9674	0.9654
5335	0.9617	0.9642	0.9659	0.9670	0.9684	0.9654

Source: The author, 2017.

Considering that each similarity score represents the quality of a candidate for the research group’s purpose, Figure 53 evidences that these 10 most *fit candidates* increased their similarity scores throughout the target interval toward the purpose of the assessment.

Figure 53 – The evolution of similarity score of the 10 most similar fit candidates for the purpose of the assessment



Source: The author, 2017.

Consequently, assessing researcher quality through their career trajectory could allow the proposed method to predict their future success. For example, the analysis of career trajectory of fit candidate number

7550, along the last five years stress the growth of productivity toward the research group's purpose.

Frame 27 shows at the left side, the most relevant accomplishments of the Fiocruz MERG, which represent the context of the group, and their respective weights (W), that is, relative relevance. The right side, Frame 27 presents the total number of accomplishments by year of the *fit candidate* 7550, who is the most similar to the research group. This similarity can be explained by increase productivity of candidate 7550 in accomplishments of great relevance in the context of the Fiocruz MERG, for instance, Journal articles with two or more coauthors and both, Membership in master and doctoral committees.

Frame 27 – The evolution of a *fit candidate* toward the research group's purpose

The purpose of Fiocruz MERG represented through weights		Fit Candidate 7550				
		Accomplishments in the target interval of career trajectory				
Accomplishment	W	2011	2012	2013	2014	2015
Masters students graduated	0.75	8	8	10	10	11
Membership in masters committees	0.74	17	17	17	18	18
Grants type a as first author	0.74	4	4	4	4	4
Doctoral students graduated concluding coadvisor	0.74	2	2	2	3	3
Membership in doctoral committees	0.73	11	12	12	14	14
Journal articles with two or more co authors	0.72	31	36	39	43	49
Journal articles as third author or other	0.71	23	28	31	34	38
Masters students graduated coadvisor	0.69	3	3	3	3	4
Journal articles as second author	0.68	7	7	7	8	10
Grants type a with two or more co authors	0.67	5	5	5	5	5
Doctoral students graduated concluding	0.64	1	2	3	3	6
Book chapters with two or more co authors	0.63	1	1	1	1	1
Book chapters as third author or other	0.57	1	1	1	1	1
Doctoral students current	0.56	0	0	0	0	4
Journal articles as first author	0.54	1	1	1	1	1
Grants type a as second author	0.53	1	1	1	1	1

Source: The author, 2017.

The quality of *fit candidate* 7550 is also evidenced in Frame 28, which shows that in 2011 and 2012, this candidate was similar to only one negative case, however, in the years 2013 to 2015, the candidate 7550 was similar to all positive cases. Consequently, to this candidate could be predict a successful research collaborator.

Frame 28 – The classification of the *fit candidate* 7550, as *fit* or *unfit*, in the last five years of career trajectory

Classification of candidate CV 7550 as <i>fit</i> or <i>unfit</i>					
Most similar case	Last five years of career trajectory				
	2011	2012	2013	2014	2015
1 <sup>st</sup> most similar	P	P	P	P	P
2 <sup>nd</sup> most similar	N	P	P	P	P
3 <sup>rd</sup> most similar	P	N	P	P	P
<b>Final classification</b>	<b>fit</b>	<b>fit</b>	<b>fit</b>	<b>fit</b>	<b>fit</b>

Source: The author, 2017.

The same individual analysis was performed considering an *unfit candidate* in some year of the target interval of the assessment. The candidate number 7859 was chosen to illustrate this example, as shown in Frame 29. This candidate has a productivity lower than the candidate 7550, for example, in the accomplishments, *Grants type a as first author*, and *Doctoral students graduated concluding coadvisor*, which are of great relevance in the context of the Fiocruz MERG, this candidate has not any accomplishment.

Frame 29 – The evolution of an *unfit candidate* in an interval of career trajectory

The purpose of Fiocruz MERG represented through weights		Unfit Candidate 7859				
		Accomplishments in the target interval of career trajectory				
Accomplishment	W	2011	2012	2013	2014	2015
Masters students graduated	0.75	7	7	7	11	11
Membership in masters committees	0.74	14	14	14	20	20
Membership in doctoral committees	0.73	23	24	25	27	32
Journal articles with two or more co authors	0.72	30	33	37	41	45
Journal articles as third author or other	0.71	26	29	33	37	41
Masters students graduated coadvisor	0.69	1	1	1	1	1
Journal articles as second author	0.68	2	2	2	2	2
Doctoral students graduated concluding	0.64	6	7	8	9	10
Journal articles as first author	0.54	8	8	8	8	8
Book chapters as first author	0.54	4	4	4	4	4
Book chapters solo authored	0.54	4	4	4	4	4
Journal articles with one co author	0.47	5	5	5	5	5
Masters students current	0.46	0	0	0	5	5
Books published with two or more co authors	0.35	0	0	0	0	1
Books published as third author or other	0.31	0	0	0	0	1
Journal articles solo authored	0.30	1	1	1	1	1
Unpublished conference with two or more co authors	0.26	1	1	1	1	1
Published conference papers as third author or other	0.16	1	1	1	1	1

Source: The author, 2017.

The quality of *candidate 7859* is also analyzed in Frame 30, which shows that in the entire period, this candidate was similar to negative cases, which can also be outliers. This can be explained due the existence of solo accomplishments, such as, *Book chapters solo authored*, and *Journal articles solo authored*. This fact could characterize the candidate 7859 as not too efficient collaborator. Furthermore, in 2014, the first similar case was a negative case, and as a consequence this candidate was considered as *unfit for the purpose*.

Frame 30 – The classification of the *candidate 7859*, as *fit* or *unfit*, in the last five years of career trajectory

<b>Classification of candidate CV 7859 as <i>fit</i> or <i>unfit</i></b>					
<b>Most similar case</b>	<b>Target interval of career trajectory</b>				
	<b>2011</b>	<b>2012</b>	<b>2013</b>	<b>2014</b>	<b>2015</b>
1 <sup>st</sup> most similar	P	P	P	N	P
2 <sup>nd</sup> most similar	N	N	P	P	N
3 <sup>rd</sup> most similar	N	P	N	N	N
<b>Final classification</b>	<b>fit</b>	<b>fit</b>	<b>fit</b>	<b>unfit</b>	<b>fit</b>

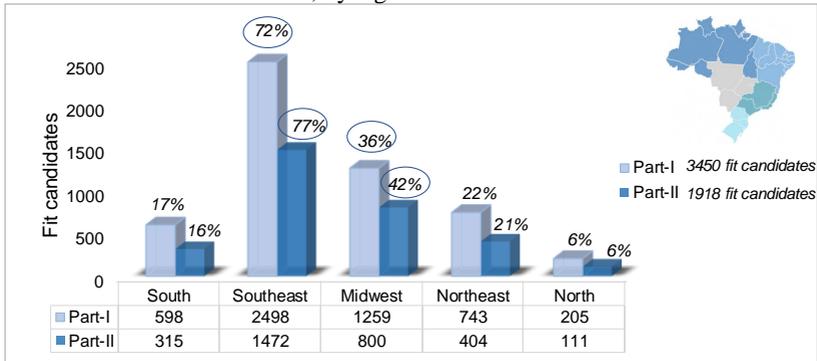
Source: The author, 2017.

Comparing these two candidates, 7550 and 7859, which although they have similar productivity, the second candidate is not so closer from the Fiocruz MERG context as the first candidate. Consequently, the candidate 7859 should not be selected as *fit for the purpose* of the assessment, and this was only realized applying the treatment of career trajectories.

At the end of this analysis, the resulting classification of *fit candidates* from the perspective of fields of study and geographic regions was present. These results were obtained, by comparing fit candidates without considering career trajectories (i.e., Part-I of experiment) and fit candidates after incorporating treatment of career trajectories (i.e., Part-II of experiment).

Figure 54 illustrates a comparison of such results from the perspective of geographic regions. It is interesting to observe the rise in the percentage of fit candidates after applying career trajectories, in the Southeast (from 72% to 77%) and Midwest (from 36% to 42%) of Brazil. There was a more concentration of fit candidates in these two regions.

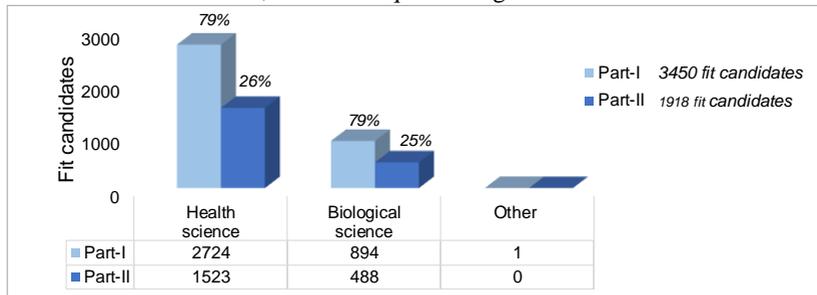
Figure 54 – The distribution of fit candidates of Experiment II, Part-I and Part-II, by regions of Brazil.



Source: The author, 2017.

Figure 55 illustrates the fit candidates of both parts of experiment II, divided by their CNPq knowledge areas. Differently from geographic regions, in this case the percentage of fit candidates remained unchanged.

Figure 55 – The distribution of fit candidates of Experiment II, Part-I and Part-II, in the CNPq knowledge areas.



Source: The author, 2017.

### 7.3.7 Concluding Remarks

In this second experiment, I demonstrated the assumption “A2.1: Incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment”. To this end, a recruitment process of candidate researchers was simulated with the purpose of selecting candidate researchers to work in collaboration with members of a research group. The experiment was demonstrated in two parts, in which the first part did not consider career

trajectory, and in the second part, the treatment of career trajectories was incorporated into the method.

*The Experiment II – Part I* dealt with data on the current production of candidate researchers. Specifically, the proposed method was able to: characterize the purpose of the assessment through weights; classify candidate researchers as fit or unfit for the purpose characterized; and rank the fit candidates in a list, to assist decision makers.

*The Experiment II – Part II* was more rigorous, including the assessment of career trajectory of candidate researchers into the proposed method, as recommended in the Leiden Manifesto (HICKS et al, 2015). Having completed this second part of the experiment, I analyzed the results and synthesize its main points.

First, the analysis compared the results of the first part of experiment with the results of the second part. Out of the 3,450 fit candidates in the first part, only 1,918 remained fit in the five consecutive years of the assessment period. This result indicated that the classification process became much stricter than the first part of the experiment.

Second, the 1,547 unfit candidates were investigated, and at least three reasons were found to explain why they are considered unfit. One reason is that some researcher did not have enough accomplishments, who may be young researchers (i.e., new researchers). Also, many candidates were classified as unfit for the purpose in some year of the target period. Finally, they were classified as fit for the purpose, but not consecutively in the entire assessment period. Thus, these results confirm that the process became more rigorous.

Third, the similarity score of the 1,918 fit candidates were investigated in each year of the target interval. This analysis allowed me to realize that the similarity scores had grown and become increasingly aligned to the purpose of the research group. Thus, these results emphasized that the treatment of career trajectory lead to better results.

Concluding, the results confirmed the assumption stated in the beginning of this section, as well as, demonstrated the alignment of the purpose-oriented method to the seventh principle of the Leiden Manifesto (HICKS et al, 2015).



## 8 CONCLUSIONS

In this chapter, I conclude this thesis, which investigated researcher quality assessment with a particular focus on research collaborators. Thus, in this conclusion I begin by answering the research question stated in chapter 1: “*How to assess researcher quality for collaborative purposes?*”. After that I address the main contributions of the proposed approach, and at the end, future works are proposed.

### 8.1 RESEARCH QUESTIONS

The research question above mentioned is supported by the sub-questions, RQ1 and RQ2, which in the following paragraphs I will describe my conclusions.

***RQ1: How to conceptualize a data model to assess researcher quality with emphasis on research collaborators?***

In order to answer RQ1, I conducted a systematic literature review, and focused on attributes that characterize research collaborators as a starting point to create a conceptual data model for data studies, as showed in Figure 24. Through this study I learned that research collaborators are primarily researchers, and their attributes originate from the relations between the researcher and the dimensions: researchers, institutions, accomplishments, and careers, as shown in chapter 4. The perception that a research collaborator is primarily a researcher suggests that the proposed model can be used for researchers in general. Furthermore, considering that the characteristics of a researcher in the collaborative or individual contexts are facets of quality, and quality depends on the purpose, the emphasis on collaboration or on any other purpose of the research assessment will be reflected by the values of the attributes of the model. For instance, the type of accomplishment attribute, whose value is coauthoring or not is considered to assess research collaborator quality.

***RQ 2: How to assess researcher quality?***

To answer RQ2, a *purpose oriented method to assess researcher quality for collaborative purposes* was proposed. This proposed method is a Knowledge Engineering (KE) approach, based on Case-Based Reasoning (CBR) methodology, and uses data from the Brazilian *Lattes*

database (lattes.cnpq.br). Furthermore, the approach is aligned with the OECD (2008), which agree with (JURAN; GODFREY, 1999) that quality is *fitness for purpose*. It is also aligned with the Leiden Manifesto principles (HICKS et al, 2015), which recommend incorporating purpose when constructing research metrics. The *purpose oriented method*, based on the CBR principle, assesses the similarity between new candidate researchers and successful researchers through their accomplishments, that is, experiences achieved by researchers along their career trajectories mapped into their CVs. Thus, the core of the proposed method is the purpose-oriented classifier, which is a CBR implementation tackled in four stages:

- i. Stage 1: Describing the problem;
- ii. Stage 2: Learning weights to represent purpose;
- iii. Stage 3: Classifying candidate researchers as fit or unfit for the purpose; and
- iv. Stage 4: Ranking candidate researchers.

Following these four stages through their proposed approaches is how the *purpose-oriented method* assess researcher quality. The detailed description of the *purpose oriented method* is found in chapter 6. The answer to this research question was supported by one hypotheses (H2.1) and one assumption (A2.1), which were demonstrated through two experiments in chapter 7.

***H2.1: A purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method***

Experiment I considered that “A *purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method*”. To teste this hypothesis, three classifiers were created, and the LOOCV (e.g., KOHAVI, 1995) was used to validate the results. These three classifiers were named, respectively, *purpose-oriented collaboration classifier* (POC1), *purpose-oriented solo classifier* (POC2), and *purpose-independent classifier* (PIC). At the end of this experiment, the PIC had correctly classified 62% of the applicants for the collaboration job, and the POC1 and POC2 had correctly classified 92% and 93%, respectively. In conclusion, these results supported the hypothesis that a purpose-oriented classifier is more accurate than a purpose independent one. Hence, this experiment demonstrated the viability of the proposed method and allowed me to continue this

investigation exploring other situations. Furthermore, it also demonstrated the alignment of the proposed method with the OECD (2008) and the second principle of the Leiden Manifesto (HICKS et al, 2015), which argues that performance should consider contextual aspects.

### **A2.1: Incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment**

Experiment II simulated a recruitment process of candidate researchers to work in collaboration with members of a research group. In order to simulate such scenario, the first challenge was to find a real example that could be used as reference in this experiment. Thus, I found in the *Microcephaly Epidemic Research Group* (MERG) ([www.cpqam.fiocruz.br/merg](http://www.cpqam.fiocruz.br/merg)), a suitable Brazilian research group undoubtedly referenced as being high-quality. This experiment was demonstrated in two parts, in which the first part did not consider career trajectory, and the second part did consider it.

Experiment II – Part I focused on data from the current production of candidate researchers, and demonstrated the *purpose-oriented method* in detail. Since the quality requirements of this application scenario contained only successful CVs, the second challenge was to solve the problem of the lack of negative instances to learn weights. Thus, the Spy technique in S-EM (LIU et al., 2002) was applied to create negative instances, and allowed me to represent the purpose of the assessment through weights. The proposed method was able to represent the purpose of the assessment through weights, and this was one of its most relevant contributions, because it allows decision makers to characterize the purpose of the assessment through the relative relevance of attributes from the set of example CVs. After that, the purpose-oriented classifier outputted 3,465 fit candidates and 11,335 unfit candidates, out of a total of 15,266 candidate researchers for the recruitment process. By analyzing the similarity among the members of the research group, an impressive result emerged revealing three distinct clusters of similarity. These clusters were the basis of the final rank that is composed of 3,450 fit candidates. Concluding this first part of experiment II, the purpose-oriented method was able to simulate a recruitment process of candidate researchers to work in collaboration with members of a research group.

Experiment II – Part II focused on the career trajectory of candidate researchers inspired by the seventh principle of the Leiden Manifesto (HICKS et al., 2015, p.430), which states that “*Even when*

*comparing large numbers of researchers, an approach that considers more information about an individual's expertise, experience, activities and influence is best*". In this second part of the experiment the same tasks performed in its first part were applied considering a target interval of the last five years of the career trajectory of candidate researchers. As a result, incorporating treatment of career trajectories into the purpose-oriented method makes it more rigorous. Thus, the results of both parts of experiment II were compared to demonstrate the assumption. For instance, out of the 3,450 fit candidates in the first part, only 1,918 remained fit in the five consecutive years of the assessment period. This result was the first indication that the classification process had become much stricter than in the first part of the experiment. In another example, the analyzes of the 1,547 unfit candidates identified three reasons why such candidates were classified as unfit for the purpose. First, some candidates were new researchers, and thus, they did not have enough accomplishments in that year. Second, some others were classified as unfit for the purpose in some of the years of the target period. Third, some were classified as fit for the purpose, but not consecutively in the entire assessment period. Finally, the 1,918 fit candidates in all of the five consecutive years of the target assessment were analyzed, and the 10 fittest were selected as sample. Observing the growth in the similarity scores for each year of the target interval allowed me to infer that their similarity scores were becoming increasingly aligned to the purpose of the research group. Thus, these three results lead me to conclude that *"incorporating treatment of career trajectories into the purpose-oriented method will produce results better aligned with the goals of the assessment"*.

Experiment II also demonstrated the alignment of the purpose-oriented method to the Leiden Manifesto principles (HICKS et al, 2015). The alignment of the proposed method to the second and seventh principles were demonstrated directly through the hypothesis H2.1 and the assumption A2.1 respectively. Nonetheless, the third principle, which emphasizes the need to acknowledge local research; and the sixth principle, which recommends taking into account the specificities of fields were considered, in section 7.3.2, in order to compose the hypothetical scenarios. At the end of this experiment, the resulting classification of fit candidates from the perspective of fields of study and geographic regions were presented. Then, I compared fit candidates from the first part of experiment with fit candidates from the second part. The result indicated that in the second part, the fit candidates were not distributed proportionally to the first part, but there was a growth in two

regions. Such results created new possibilities for future studies on assessment purposes that require treatment of fields of study and geographic regions.

In conclusion, I look back to the main research question, “*How to assess researcher quality for collaborative purposes?*”, to answer that this thesis proposed an appropriate method to assess both, researcher’s quality and research collaborator’s quality. This thesis investigated researchers in general, however it focused particularly on the research collaborator as a unit of analysis in the assessment processes, such as, selection, hiring, and funding decisions. In sum, this research question was answered through the experimental collaborative scenarios and simulated processes of recruitment of research collaborators, by applying the *purpose-oriented method*.

## 8.2 CONTRIBUTIONS

In this section I present the potential contributions of this thesis, which are organized in order to emphasize the proposed method, and the practical contributions to decisions in science and technology.

This thesis proposes a *purpose-oriented method* which different from the methods based on coauthorship networks that consider the entire network of collaborators, the proposed method focuses exclusively on the individual research collaborator as unity of analysis. Moreover, this proposed method is based on the researchers’ career trajectory, through their *curriculum vitae*, which can contain accomplishments in collaborative purposes that are not commonly considered in metrics based on citation index. For example, *the total number of doctoral committees that a researcher is member*. Thus, the *purpose-oriented method* considers such accomplishments, and hence, contributes with a novel approach to assess research collaborators, which as recommended by Bozeman, Fay e Slade (2013), goes beyond the citation index.

This study adds knowledge to the fields of Knowledge Engineering, and Bibliometrics. In this perspective, it contributes by extending the literature in research collaboration by enhancing the comprehension of the domain knowledge associated with research collaborators. For instance, it contributes with a data model that represents the domain knowledge about research collaborators; It categorizes a set of types of accomplishments as solo or collaborative purpose; and it uses knowledge engineering methodology, such as the Case-based reasoning to propose a method based on career trajectory to Bibliometrics fields.

Another contribution equally important stresses the correct use of research metrics, in that, it demonstrates the relevance of the principles stated in the Leiden Manifesto (HICKS et al, 2015). For example, the first principle states that *quantitative evaluation should support qualitative expert assessment*. The proposed method compares automatically large numbers of researcher CVs, in an open, transparent and simple process, which provides decision makers with objective and consistent information to guide their judgments. The second principle is emphasized by the proposed method in Experiment I, which considers that *a purpose-oriented method to assess researcher quality is more accurate than a purpose-independent method*. The third principle recommends that locally relevant research should be protected, and this is assumed by the *purpose-oriented method*, in that, it considers the contextual aspects of assessment, such as, fields of study, geographical regions, and types of accomplishments. The seventh principle inspires Experiment II, in that, *the treatment of career trajectory is considered in the assessment process*. Consequently, the alignment of the *purpose-oriented method* with the principles of the Manifesto (HICKS et al, 2015), leads its recommendation in assessments concerning collaborative purposes.

Of course, from the perspective of science and technology, there is a number of practical contributions of great importance for decision makers. The two experiments presented in this study reflect the usefulness of the *purpose-oriented method* in selection processes of research collaborators. The first experiment, for instance, shows that the proposed method contributes with the assessment of researchers in general, in both, individual and collaborative purposes. The second experiment presents the *purpose-oriented method* not only as a method, but also a methodology that describes step-by-step how to assess research collaborators based on their career trajectories. Next, more two contributions demonstrated in the second experiment will be highlighted.

One of the most significant contributions of the proposed method is its capability of representing the criteria of the assessment through the relative relevance of the attributes of successful researchers. This fact allows decision makers to enhance the comprehension about the collaborative purpose of the assessment. For example, analyzing the set of example CVs makes possible to understand different clusters of researchers, as well as, the role of individual researchers in collaborative purposes.

The other specific contribution of the proposed method concerns the possibility of assessing researchers and research collaborators along their career trajectories. I believe that this fact opens a myriad of options

for decision makers. For example, they could trace the career trajectories, by contrasting the experiences of candidate researchers with successful researchers. This could give to them a more analytical focus on the assessments of individuals and groups along the time, making possible to predict their trends in collaborations.

I conclude this section, with my particular view about the contributions of this study. The *purpose-oriented method* is an instrument of governance for the science in collaboration, which assists decision makers to promote a transparent and more efficient management of research collaboration. For example, considering funding agencies, the purpose-oriented method could help in the evaluation of research projects for collaborative purposes, in which the evaluation of team is commonly a complex task based on subjective criteria analyzed by peer-review committees. The *purpose-oriented method* is also a suitable instrument for universities, contributing with the management and evaluation of research groups. Finally, I believe that the proposed method has much to contribute with the emergency of digital technologies that support a new culture of collaborative awareness.

### 8.3 FUTURE WORKS

In addition to contributions above mentioned, some directions for future works are suggested in the following paragraphs.

An interesting future work is directed to the range of possibilities offered by the study of career trajectories. For instance, maturity cycle could be incorporated to the *purpose-oriented method*, in order to trace the evolution of the career stages that a research collaborator achieve. In order to base the study of career trajectories, the investigation of CBR Time-Series Analysis, and CBR Explanation could be essential requirements.

Furthermore, the combination between Case-Based Reasoning (CBR) and Social Network Analysis (SNA) could be another interesting field study to explore. This link could provide significant results in the proposing of new models and methods to assess research collaborators.

Concerning the *Conceptual data model for research collaborators*, an interesting future work could be the possibility of exploring alternative representational formalisms based on ontologies. For example, the representation of the data model in a *first-order-ontology*.

Other essential studies are related to the advancements in the *purpose-oriented method*, for example to improve the sensibility of

strategies to select likely negative instances that could be more generalizable.

Concluding the suggestions for future works, a notable and fascinating challenge, which transcend this thesis, is the study of Coproduction. The wealth of coproduction opens innumerable possibilities of future works, extending the proposed method beyond its application in research collaboration. For example, the *purpose-oriented method* could be simulated in environments of innovation and citizenscience, with different types of CV databases, combining data from universities, funding agencies, research institutes, business, and government. Consequently, the *purpose-oriented method* could be packed as a product for application in large scale.

## REFERENCES

- AAMODT, A.; PLAZA, E. Case-Based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, v. 7, n. 1, p. 39–59, 1994.
- ABOWD, G. et al. Towards a better understanding of context and context-awareness. In: *Handheld and ubiquitous computing*. Springer Berlin/Heidelberg, 1999. p. 304–307.
- ABRAMO, G.; D'ANGELO, C. A. National-scale research performance assessment at the individual level. *Scientometrics*, v. 86, n. 2, p. 347–364, 2011.
- ABRAMO, G.; D'ANGELO, C. A. A farewell to the MNCS and like size-independent indicators: Rejoinder. *Journal of Informetrics*, v. 10, p. 679–683, 2016.
- ABRAMO, G.; D'ANGELO, C. A.; DI COSTA, F. Research collaboration and productivity: Is there correlation? *Higher Education*, v. 57, n. 2, p. 155–171, 2009.
- ABRAMO, G.; D'ANGELO, C. A.; DI COSTA, F. Investigating returns to scope of research fields in universities. *Higher Education*, v.68, n.1, p.69-85, 2014.
- ABRAMO, G.; D'ANGELO, C. A.; MURGIA, G. The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*, v. 7, n. 2, p. 442–454, 2013a.
- ABRAMO, G.; D'ANGELO, C. A.; MURGIA, G. Gender differences in research collaboration. *Journal of Informetrics*, v. 7, n. 4, p. 811–822, 2013b.
- ABRAMO, G.; D'ANGELO, C. A.; MURGIA, G. Variation in research collaboration patterns across academic ranks. *Scientometrics*, v. 98, n. 3, p. 2275–2294, 2014.
- ABRAMO, G.; D'ANGELO, C. A.; ROSATI, F. A methodology to measure the effectiveness of academic recruitment and turnover. *Journal of Informetrics*, v. 10, n. 1, p. 31-42, 2016.
- ADAMS, J. Collaborations: the fourth age of research. *Nature*, v. 497, n. 7451, p. 557–560, 2013.
- ADAMS, J. Supplementary data to: The Fourth Age of Research. Comment in. *Nature*, p. 557–560, 2014.

AGORGIANITIS, I. et al. Evaluating distributed methods for CBR systems for monitoring business process workflows. In: ICCBR Workshops, p. 122–131, 2016.

AHA, D. Feature weighting for lazy learning algorithms. *Feature Extraction, Construction and Selection*, p. 1–20, 1998.

AHA, D. W. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, v. 36, n. 2, p. 267–287, 1992.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Machine Learning*, v. 6, p. 37–66, 1991.

AHMED, M. U.; BEGUM, S.; FUNK, P. Case Studies on the Clinical Applications using Case-Based Reasoning. In: *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*. IEEE, p.3–10, 2012.

AJIFERUKE, I.; BURELL, Q.; TAGUE, J. Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics*, v. 14, n. 5–6, p. 421–433, 1988.

ALLIK, J. Factors affecting bibliometric indicators of scientific quality. *Trames*, v. 17, n. 3, p. 199–214, 2013.

AMERICAN SOCIETY FOR CELL BIOLOGY. San Francisco Declaration on Research Assessment: Putting science into the assessment of research. 2015.

ARMITAGE, A.; KEEBLE-ALLEN, D. Undertaking a structured literature review or structuring a literature review: Tales from the field. *Electronic Journal of Business Research Methods*, v. 6, n. 2, p. 103–114, 2008.

ARTHUR, M. B.; HALL, D. T.; LAWRENCE, B. S. Generating new directions in career theory: The case for a transdisciplinary approach. In: *Handbook of career theory*. Cambridge University Press, p. 7–25, 1989.

ATKINS, D. E. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. 2003.

BALANCIERI, R. Um método baseado em ontologias para explicitação de conhecimento derivado da análise de redes sociais de um domínio de aplicação. PhD thesis, 2010.

BEAVER, D. DEB. Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, v. 52, n. 3, p. 365–377, 2001.

- BEAVER, D. DEB. Does collaborative research have greater epistemic authority? *Scientometrics*, v. 60, n. 3, p. 399–408, 2004.
- BEAVER, D. DEB. Quantity is only one of the qualities. *Scientometrics*, v. 93, n. 1, p. 33–39, 2012.
- BEAVER, D. DEB.; ROSEN, R. Studies in scientific collaboration: Part I. The Professional Origins of Scientific Co-Authorship. *Scientometrics*, v. 1, n. 1, p. 65–84, 1978.
- BEAVER, D. DEB.; ROSEN, R. Studies in scientific collaboration: Part II. Scientific co-authorship, research productivity and visibility in the french scientific elite, 1799-1830. *Scientometrics*, v. 1, n. 2, p. 133–149, 1979a.
- BEAVER, D. DEB.; ROSEN, R. Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, v. 1, n. 3, p. 231–245, 1979b.
- BECK, M. T. Editorial Statements *Scientometrics*, 1978.
- BELLOTTI, E.; KRONEGGER, L.; GUADALUPI, L. The evolution of research collaboration within and across disciplines in Italian Academia. *Scientometrics*, v. 109, n. 2, p. 783–811, 2016.
- BERGER, P. L.; LUCKMANN, T. *The social construction of reality: A treatise in the sociology of knowledge*. Penguin UK, 1991.
- BERGSTROM, C. T. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, v. 68, p. 314–316, 2007.
- BERNELA, B.; MILARD, B. *Co-authorship Network Dynamics and Geographical Trajectories - What Part Does Mobility Play?* SAGE Publications, v. 131, n. 1, p. 5–24, 2016.
- BHATTACHERJEE, A. *Social Science Research: principles, methods, and practices*. 2012.
- BOARDMAN, P. C.; CORLEY, E. A. University research centers and the composition of research collaborations. *Research Policy*, v. 37, n. 5, p. 900–913, 2008.
- BORDIN, A. S. *Framework baseado em conhecimento para análise de rede de colaboração científica*. PhD thesis, 2015.
- BORNMANN, L.; HAUNSCHILD, R. Relative Citation Ratio (RCR): A first empirical attempt to study a new field-normalized bibliometric indicator. *Journal*, p. 2015–11, 2015.

BORNMANN, L.; HAUNSCHILD, R. To what extent does the Leiden Manifesto also apply to altmetrics? A discussion of the manifesto against the background of research into altmetrics. *Online Information Review*, v. 40, n. 4, p. 529–543, 2016.

BORROR, D. J. Dictionary of word roots and combining forms. Compiled from the Greek, Latin, and other languages, with special reference to biological terms and scientific names. 1960.

BOVO, A. B. Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais. PhD thesis, 2011.

BOZEMAN, B.; CORLEY, E. A. Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy*, v. 33, n. 4, p. 599–616, 2004.

BOZEMAN, B.; FAY, D.; SLADE, C. P. Research collaboration in universities and academic entrepreneurship: The-state-of-the-art. *Journal of Technology Transfer*, v. 38, n. 1, p. 1–67, 2013.

BOZEMAN, B.; GAUGHAN, M.; YOUTIE, J. Social dynamics of research collaboration: Norms, practices, and ethical issues in determining coauthorship rights. *Scientometrics*, v. 101, n. 2, p. 953–962, 2014.

BRAGLIA, I. DE A. Um Modelo Baseado em Ontologia e Extração de Informação como Suporte ao Processo de Design Instrucional na Geração de Mídias do Conhecimento. PhD thesis, 2014.

BREMBS, B. .; BUTTON, K. .; MUNAFÒ, M. . Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, n. JUN, 2013.

BROCKE, J. VOM; LIPPE, S. Identifying and Managing Creative Tasks in Collaborative IS Research Projects. *Project Management Journal*, v. 44, n. 6, p. 94–113, 2013.

BROWN, C. G. Rethinking the fundamentals of international scientific cooperation in the early twenty-first century. In: *Information Technology Applications in Biomedicine, 2000. Proceedings. 2000 IEEE EMBS International Conference on. IEEE, 2000. p. 136-145.*

BRUDNEY, J.; ENGLAND, R. Toward a definition of the coproduction concept. *Public Administration Review*, v. 43, n. 1, p. 59–65, 1983.

BRUMBACK, R. A. “3.. 2.. 1.. Impact [factor]: Target [academic career] destroyed!?”: Just another statistical casualty. *Journal of Child Neurology*, v. 27, n. 12, p. 1565–1576, 2012.

- BUTLER, D. Zika Detective: In 366 Days: the year in science: Nature's 10. Nature, v. 540, p. 22–29, 2016.
- CAÑIBANO, C.; BOZEMAN, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. Research Evaluation, v. 18, n. 2, p. 86–94, 1 jun. 2009.
- CARBONELL, J. G.; MICHALSKI, R. S.; MITCHELL, T. M. Machine Learning: A Historical and Methodological Analysis. AI Magazine, v. 4, n. 3, p. 69, 1983.
- CARILLO, M. R.; PAPAGNI, E.; SAPIO, A. Do collaborations enhance the high-quality output of scientific institutions? Evidence from the Italian Research Assessment Exercise. Journal of Socio-Economics, v. 47, p. 25–36, 2013.
- CARO, L. DI; CATALDI, M.; SCHIFANELLA, C. The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. Scientometrics, v. 93, p. 583–607, 2012.
- CASTELVECCHI, D. Physics paper sets record with more than 5,000 authors. Nature, 2015.
- CECI, F. Um modelo baseado em casos e ontologia para apoio à tarefa intensiva em conhecimento de classificação com foco na análise de sentimentos. PhD thesis, 2015.
- CGEE. Mestres e doutores 2015 - Estudos da demografia da base técnico-científica brasileira. Brasília, DF: Centro de Gestão e Estudos Estratégicos, 2016.
- CHIVA, R. Organizational Learning and Organizational Knowledge: Towards the Integration of Two Approaches. Management Learning, v. 36, n. 1, p. 49–68, 2005.
- CHOMPALOV, I.; GENUTH, J.; SHRUM, W. The organization of scientific collaborations. Research Policy, v. 31, n. 5, p. 749–767, 2002.
- CMS COLLABORATION. The CMS experiment at the CERN LHC. JINST, v. 3, p. 285, 2008.
- CNPQ. National Council for Scientific and Technological Development. Available in: <<http://lattes.cnpq.br/>>. Accessed in: 13<sup>rd</sup> Jul. 2017.
- COLLINS, F. S.; MORGAN, M.; PATRINOS, A. The Human Genome Project: Lessons from Large-Scale Biology. Science, v. 300, n. 5617, p. 286–290, 2003.
- COLLINS, J. P. Sailing on an Ocean of 0s and 1s. Science. v. 327p. 1455–1456, 2010.

COOMBS, S. K.; PETERS, I. The Leiden Manifesto Under Review : What Libraries Can Learn From It. *Digital Library Perspectives*, p. 1–14, 2017.

CORLEY, E. A.; BOARDMAN, P. C.; BOZEMAN, B. Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research Policy*, v. 35, n. 7, p. 975–993, 2006.

CORLEY, E. A.; SABHARWAL, M. Scholarly collaboration and productivity patterns in public administration: Analysing recent trends. *Public Administration*, v. 88, n. 3, p. 627–648, 2010.

CRESWELL, J. W. *Research design: qualitative, quantitative, and mixed methods approaches*. 3<sup>rd</sup> ed. Sage publications, 2009.

CUMMINGS, J. N.; KIESLER, S. Collaborative Research Across Disciplinary and Organizational Boundaries. *Social Studies of Science*, v. 35, n. 5, p. 703–722, 2005.

CUMMINGS, J. N.; KIESLER, S. Coordination costs and project outcomes in multi-university collaborations. *Research Policy*, v. 36, n. 10, p. 1620–1634, 2007.

CUMMINGS, J. N.; KIESLER, S. Organization theory and new ways of working in science. In: *Science and Innovation Policy, 2011 Atlanta Conference on*. IEEE, p. 1-5, 2011.

CUMMINGS, J. N.; KIESLER, S. Organization Theory and the Changing Nature of Science. *Journal of Organization Design*, v. 3, n. 3, p. 1–16, 2014.

DAHLANDER, L.; MCFARLAND, D. Ties That Last: Tie Formation and Persistence in Research Collaborations over Time. *Administrative Science Quarterly*, v. 58, p. 69–110, 2013.

DAVENPORT, T. H.; PRUSAK, L. *Working Knowledge How Organization Manage What They Know*. Harvard Business School Press, n. January 1998, p. 1–15, 1998.

DAVID, D.; FRANGOPOL, P. The lost paradise, the original sin, and the Dodo bird: a scientometrics Sapere Aude manifesto as a reply to the Leiden manifesto on scientometrics. *Scientometrics*, v. 105, n. 3, p. 2255–2257, 2015.

DE STEFANO, D. et al. The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, v. 35, n. 3, p. 370–381, 2013.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, v. 39, n. 1, p. 1–38, 1977.

DÉTIENNE, F.; BAKER, M.; BURKHARDT, J.-M. Quality of collaboration in design meetings: Methodological reflexions. *CoDesign*, v. 8, n. 4, p. 247–261, 2012.

DGP.CNPQ. Grupo de Pesquisa da Epidemia de Microcefalia (MERG). Available in: <<http://dgp.cnpq.br/dgp/espelhogrupo/2723404431935999>>. Accessed in: 11<sup>st</sup> Jun. 2017.

DICKINS, T. E. Social Constructionism as Cognitive Science. *Journal for the Theory of Social Behaviour*, v. 34, n. 4, p. 333–352, 2004.

DIDEGAH, F.; THELWALL, M. Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, v. 7, n. 4, p. 861–873, 2013.

DIETZ, J. S. et al. Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. *Scientometrics*, v. 49, n. 3, p. 419–442, 2000.

DUARTE, J.C.B. *Poemas de Júlio Carlos Duarte*, 1990.

DUARTE, K.B.; WEBER, R. O.; PACHECO, R.C.S. Conceptual data model for research collaborators. In: VI International Conference on Knowledge and Innovation (CIKI 2016). 2016a.

DUARTE, K.; WEBER, R.; PACHECO, R. C. S. Purpose-oriented metrics to assess researcher quality. In: 21<sup>st</sup> International Conference on Science and Technology Indicators (STI2016): Peripheries, frontiers and beyond, p.1312-1314, 2016b.

DUARTE, K.B.; WEBER, R.O.; PACHECO, R.C.S. Case-Based Comparison of Career Trajectories. In: CEUR Workshop Proceedings, 24th International Conference on Case-Based Reasoning Workshops (ICBR-WS 2016), v.1815, p.152-161, 2016c.

EL-SAPPAGH, S. H.; ELMOGY, M. Case Based Reasoning: Case Representation Methodologies. *International Journal of Advanced Computer Science and Applications*,(IJACSA), v. 6, n. 11, p. 192–208, 2015.

ELSEVIER. Scopus: Content Coverage Guide. Available in: <[https://www.elsevier.com/\\_data/assets/pdf\\_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf](https://www.elsevier.com/_data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf)>. Accessed in: 16<sup>th</sup> Jun. 2014.

ENENGEL, B. et al. Co-production of knowledge in transdisciplinary doctoral theses on landscape development-An analysis of actor roles and knowledge types in different research phases. *Landscape and Urban Planning*, v. 105, n. 1–2, p. 106–117, 2012.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD-96*, v.96, n.34, p.226-231, 1996.

ETZKORN, B. Data normalization and standartization, 2011. Available in: <<http://www.benetz Korn.com/2011/11/data-normalization-and-standardization/>>. Accessed in: 21<sup>st</sup> Aug. 2016.

EVANS, J. A; FOSTER, J. G. Metaknowledge. *Science*, v. 331, n. 6018, p. 721–725, 2011.

GIORE, S. M. Interdisciplinarity as teamwork: How the science of teams can inform team science. *Small Group Research*, v. 39, n. 3, p. 251–277, 2008.

FORTE, A.; LAMPE, C. Defining, Understanding, and Supporting Open Collaboration: Lessons From the Literature. *American Behavioral Scientist*, v. 57, n. 5, p. 535–547, 2013.

FRIEDEN, T.; Celina Turchi. In: *TIME 100: The 100 Most Influential People in the World*. TIME.com. 2017. Available in: <<http://time.com/collection/2017-time-100/4742680/celina-turchi/>>. Accessed in: 20 Apr 2017.

FRIEDMAN, M.; MING, M.; KANDEL, A. On the Theory of Typicality. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, v. 3, n. 2, p. 127–142, 1995.

FRIEDMAN, T. *The world is flat: A brief history of the twenty-first century*. Macmillan, 2005.

FURTADO, C. A. et al. A spatiotemporal analysis of Brazilian science from the perspective of researchers' career trajectories. *PLoS ONE*, v. 10, n. 10, p. 1–28, 2015.

GADD, E. Looking for Leiden : Let's Make Use of ALL Available Metrics. Available in: <<http://www.socialsciencespace.com/2015/11/looking-for-leiden-lets-make-use-of-all-available-metrics/>>. Accessed in: 15<sup>th</sup> Jul. 2016.

GALYAVIEVA, M. S. On the Formation of the Concept of Informetrics (Review). *Scientific and Technical Information Processing*, v. 40, n. 2, p. 89–96, 2013.

GARFIELD, E. Citation indexes for science; a new dimension in documentation through association of ideas. *Science*, v. 122, n. 122, p. 108–11, 1955.

- GARFIELD, E. Science Citation Index. An International Interdisciplinary Index to the Literature of Science. Institute for Scientific Information, v. 1, p. 649–654, 1963.
- GARFIELD, E. “Science Citation Index” - A New Dimension in Indexing. *Science*, v. 144, n. 3619, p. 649–654, 1964.
- GARFIELD, E. The evolution of the science citation index. *International Microbiology*, v. 10, n. 1, p. 65–69, 2007.
- GARFIELD, E.; MALIN, M. Can Nobel Prize winners be predicted. 135th Annual Meeting, American Association for the Advancement of Science. Anais...Dallas, Texas: 1968
- GAUGHAN, M. Using the curriculum vitae for policy research: an evaluation of National Institutes of Health center and training support on career trajectories. *Research Evaluation*, v. 18, n. June, p. 117–124, 2009.
- GAUGHAN, M.; BOZEMAN, B. Center funding. *Research Evaluation*, v. 11, n. 1, p. 17–26, 2002.
- GAZNI, A.; DIDEGAH, F. Investigating different types of research collaboration and citation impact: A case study of Harvard University’s publications. *Scientometrics*, v. 87, n. 2, p. 251–265, 2011.
- GAZNI, A.; SUGIMOTO, C. R.; DIDEGAH, F. Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, v. 63, n. 2, p. 323–335, 2012.
- GAZNI, A.; THELWALL, M. The long-term influence of collaboration on citation patterns. *Research Evaluation*, v. 23, n. 3, p. 261–271, 2014.
- GIGCH, J. P. A methodological comparison of the science, systems and metasystem paradigms. *International Journal of Man-Machine Studies*, v. 11, n. 5, p. 651–663, 1979.
- GROUP, M. E. R. Microcephaly in Infants, Pernambuco State, Brazil, 2015. *Emerging infectious diseases*, v. 22, n. 6, p. 1090–1093, 2016.
- GU, M.; AAMODT, A. Evaluating CBR systems using different data sources: A case study. *Advances in Case-Based Reasoning*, p. 121–135, 2006.
- GUNAWARDENA, S. Recommending Research Profiles for Multidisciplinary Academic Collaboration. Thesis 2013.
- GUNAWARDENA, S.; WEBER, R. O.; STOYANOVICH, J. Learning feature weights from positive cases. In: *International Conference on Case-Based Reasoning*. Springer, Berlin, Heidelberg, v. 7969, p. 134–148, 2013.

GUNDERSEN, O. E. Enhancing the Situation Awareness of Decision Makers by Applying Case-Based Reasoning on Streaming Data. PhD thesis, 2014.

HALL, M. A. Correlation-based Feature Selection for Machine Learning. PhD thesis, 1999.

HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for data mining. *IEEE Transactions on Knowledge and Data Engineering*, v. 15, n. 3, p. 1–16, 2003.

HARZING, A. W.; ALAKANGAS, S. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, v. 106, n. 2, p. 787–804, 2016.

HAUSTEIN, S.; LARIVIÈRE, V. The use of bibliometrics for assessing research: Possibilities, limitations and adverse effects. *Incentives and Performance: Governance of Research Organizations*, p. 1–14, 2015.

HENNEMANN, S.; RYBSKI, D.; LIEFNER, I. The myth of global science collaboration-Collaboration patterns in epistemic communities. *Journal of Informetrics*, v. 6, n. 2, p. 217–225, 2012.

HEY, T.; HEY, J. e-Science and its implications for the library community. *Library Hi Tech*, v. 24, n. 4, p. 515–528, 2006.

HICKS, D. et al. The Leiden Manifesto for research metrics. *Nature*, v. 520, n. 7548, p. 9–11, 2015.

HICKS, D.; MELKERS, J. Bibliometrics as a Tool for Research Evaluation. *Handbook on the theory and practice of program evaluation*, p. 323-249, 2012.

HILL, S. J. et al. BIG science: A collaborative framework for large scale research. In: *Proceedings of the International 18<sup>th</sup> International Conference on Supporting Group Work*. ACM, p. 285-287, 2014.

HIRSCH, J. E. An index to quantify an individual's scientific research output. *PNAS*, v. 102, n. 46, p. 16569–16572, 2005.

HOPGOOD, A. A. The State of Artificial Intelligence. *Advances in Computers*, v. 65, n. December 2005, p. 1–75, 2005.

HULME, E. W. *Statistical bibliography in relation to the growth of modern civilization*. London: Butler & Tanner. 1923.

HUNTER, L.; LEAHEY, E. Collaborative research in sociology: Trends and contributing factors. *American Sociologist*, v. 39, n. 4, p. 290–306, 2008.

- IBÁÑEZ, A.; BIELZA, C.; LARRAÑAGA, P. Relationship among research collaboration, number of documents and number of citations: A case study in Spanish computer science production in 2000-2009. *Scientometrics*, v. 95, p. 689–716, 2013.
- IENCO, D.; PENSA, R. G. Positive and unlabeled learning in categorical data. *Neurocomputing*, v. 196, p. 113–124, 2016.
- JAYALAKSHMI, T.; SANTHAKUMARAN, A. Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*, v. 3, n. 1, p. 1–5, 2011.
- JEONG, S.; CHOI, J. Y.; KIM, J.-Y. On the drivers of international collaboration: The impact of informal communication, motivation, and research resources. *Science and Public Policy*, v. 41, n. 4, p. 520–531, 2014.
- JOHN, G. H. G.; LANGLEY, P. Estimating Continuous Distributions in Bayesian Classifiers. In: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 338-345.1995
- JONKERS, K.; CRUZ-CASTRO, L. Research upon return: The effect of international mobility on scientific ties, production and impact. *Research Policy*, v. 42, n. 8, p. 1366–1377, 2013.
- JONKERS, K.; TIJSSSEN, R. Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics*, v. 77, n. 2, p. 309–333, 2008.
- JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, v. 349, n. 6245, p. 255–260, 2015.
- JURAN, J. M.; GODFREY, A. B. *Juran's Quality Handbook*. 5<sup>th</sup> edition, McGraw-Hill, 1999.
- KATZ, J. S.; MARTIN, B. R. What is research collaboration? *Research Policy*, v. 26, n. 1, p. 1–18, mar. 1997.
- KIRA, K.; RENDELL, L. The feature selection problem: Traditional methods and a new algorithm. *Aaai*, p. 129–134, 1992.
- KLEIN, J. T. Evaluation of Interdisciplinary and Transdisciplinary Research. v. 35, p. 116–123, 2008.
- KLEIN, J. T. Prospects for transdisciplinarity. **Futures**, v. 36, n. 4, p. 515–526, 2004.
- KNOBEL, M.; PATRICIA SIMÕES, T.; DE BRITO CRUZ, C. International collaborations between research universities: Experiences and best practices. *Studies in Higher Education*, v. 38, n. 3, p. 405–424, 2013.

KOCSIS, T. et al. Case-Based Reasoning system for mathematical modelling options and resolution methods for production scheduling problems: Case representation, acquisition and retrieval. *Computers and Industrial Engineering*, v. 77, p. 46–64, 2014.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, v. 14, n. 12, p. 1137–1143, 1995.

KONONENKO, I.; ROBNIK-SIKONJA, M.; POMPE, U. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. *Artificial intelligence: methodology, systems, applications*, p. 31–40, 1996.

KUHN, T. S. *The structure of scientific revolutions*. 1st ed. ed. [s.l.] University of Chicago Press, 1962.

KUMAR, S. Co-authorship networks: A review of the literature. *Aslib Journal of Information Management*, v. 67, n. 1, p. 55–73, 2015.

KUMAR, S.; JAN, J. M. Mapping research collaborations in the business and management field in Malaysia, 1980-2010. *Scientometrics*, v. 97, p. 491–517, 2013.

KURBALIJA, V. *Time series analysis and prediction using case based reasoning technology*. PhD thesis, 2009.

LANE, J. Let's make science metrics more scientific. *Nature*, v. 464, n. 7288, p. 488–9, 25 mar. 2010.

LANE, J. I. *Measuring Science: Bibliometrics and Beyond*. *Art and Science of Science and Technology: Proceedings of the Forum and Roundtable*. Anais...Albuquerque, New Mexico: Science, Technology, and Public Policy Program. Harvard Kennedy School, 2013

LANE, J. I.; OWEN-SMITH, J.; ROSEN, R. F.; WEINBERG, B. A. New linked data on research investments: Scientific workforce, productivity, and public value. *Research Policy*, v. 44, n. 9, p. 1659–1671, 2015.

LARGENT, M. A.; LANE, J. I. Star Metrics and the Science of Science Policy. *Review of Policy Research*, v. 29, n. 3, p. 431–438, 2012.

LAWANI, S. M.; ROAD, O. Some Bibliometric Correlates of Quality in Scientific Research. *Scientometrics*, v. 9, n. 1–2, p. 13–25, 1986.

LEAHEY, E. From Sole Investigator to Team Scientist: Trends in the Practice and Study of Research Collaboration. *Annual Review of Sociology*, v. 42, n. 1, p. 81–100, 2016.

- LEAHEY, E.; REIKOWSKY, R. C. Research specialization and collaboration patterns in sociology. *Social Studies of Science*, v. 38, n. 3, p. 425–440, 2008.
- LEE, S.; BOZEMAN, B. The impact of research collaboration on scientific productivity. *Social Studies of Science*, v. 35, n. 5, p. 673–702, 2005.
- LEIDEN MANIFESTO BLOG. Leiden Manifesto wins 2016 EASST Ziman award. In: *Leiden Manifesto for Research Metrics (Blog)*. Available in: <<http://www.leidenmanifesto.org/blog/leiden-manifesto-wins-2016-easst-ziman-award>>. Accessed in: 17<sup>th</sup> Jul. 2017
- LEYDESDORFF, L. Theories of citation? *Scientometrics*, v. 43, n. 1, p. 5–25, 1998.
- LEWIS-BECK, M.; BRYMAN, A. E.; LIAO, T. F. *The Sage encyclopedia of social science research methods*. Sage Publications, 2004.
- LIU, B. et al. Partially Supervised Classification of Text Documents. In: *ICML 2002 - Proceedings of the 9<sup>th</sup> International Conference on Machine Learning*. v.2, p.387-394, 2002.
- LIU, B. et al. Building text classifiers using positive and unlabeled examples. In: *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003.
- MAGLAUGHLIN, K. L.; SONNENWALD, D. H. Factors that impact interdisciplinary natural science research collaboration in academia. In: *International Society for Scientometrics and Informetrics (ISSI)*, p.1-12, 2005.
- MAKKAR, K. A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets. *IJERMT*, v. 4, n. 6, p. 195–198, 2015.
- MANHÃES, M. C. Innovativeness and prejudice: designing a landscape of diversity for knowledge creation. PhD Thesis, 2015.
- MÁNTARAS, L. M. et al. Retrieval, reuse, revision, and retention in case-based reasoning. In: *The Knowledge Engineering Review*. [s.l.] Cambridge University Press, 2005. v. 0, p. 1–24.
- MARK, W. S.; SIMPSON, R. L. *Knowledge-Based Systems an Overview*. *IEEE Expert*, v. 63, n. 3, p. 12–17, 1991.
- MARQUES, J. S. Reforming Technology Company Incentive Programs for Achieving Knowledge-Based Economic Development: a Brazil-Australia comparative study Florianópolis. . PhD Thesis, 2016.

MARZOLLA, M. Assessing evaluation procedures for individual researchers: The case of the Italian National Scientific Qualification. *Journal of Informetrics*, v. 10, n. 2, p. 408–438, 2016.

MASON, K. Participatory Action Research: Coproduction, Governance and Care. *Geography Compass*, v. 9, n. 9, p. 497–507, 2015.

MAUTHNER, N. S.; DOUCET, A. “Knowledge once divided can be hard to put together again”: An epistemological critique of collaborative and team-based research practices. *Sociology*, v. 42, n. 5, p. 971–985, 2008.

MCDOUGALL-WATERS, J. et al. *Philosophical Transactions : 350 years of publishing at the Royal Society (1665–2015)*, The Royal Society, 2015.

MEIJER, A. Co-production in an Information Age: Individual and Community Engagement Supported by New Media. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, v. 23, n. 4, p. 1156–1172, 18 jul. 2012.

MERG. Grupo de Pesquisa da Epidemia de Microcefalia. Apresentação. Available in: <<http://www.cpqam.fiocruz.br/merg/>>. Accessed in: 11 jun. 2017.

MILLER, G. A. The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, v. 7, n. 3, p. 141–144, 2003.

MILLER, T. R. et al. Epistemological Pluralism: Reorganizing Interdisciplinary Research. *Ecology and Society*, v. 13, n. 2, p. 46, 2008.

MINGERS, J.; LEYDESDORFF, L. A review of theory and practice in scientometrics. *European Journal of Operational Research*, v. 246, n. 1, p. 1–19, 2015.

MINSKY, M. Steps Toward Artificial Intelligence. *Proceedings of the Ire*, p. 8–30, 1961.

MOED, H. Research Assessment 101: An introduction – Research Assessment. *Research Trends*, n. 23, 2011.

MOLAS-GALLART, J. Research Governance and the Role of Evaluation: A Comparative Study. *American Journal of Evaluation*, p. 1–16, 2012.

MOODY, J. The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. v. 69, 2004.

NACKE, O. Informetrie: Ein neuer name für eine neue disziplin. *Nachrichten für Dokumentation*, v. 30, n. 6, p. 219–226, 1979.

NALIMOV, V. V. Quantitative methods in the study of the process of scientific development. *Voprosy Filosofii*, v. 20, n. 12, p. 38–47, 1971.

- NARIN, F.; HAMILTON, K. S. Bibliometric performance measures. *Scientometrics*, v. 36, n. 3, p. 293–310, 1996.
- NARIN, F.; OLIVASTRO, D.; STEVENS, K. A. Bibliometrics/Theory, Practice and Problems. *Evaluation Review*, v. 18, n. 1, p. 65–76, 1994.
- NAZÁRIO, D. C.; DANTAS, M. A. R.; TODESCO, J. L. Knowledge engineering: Survey of methodologies, techniques and tools. *IEEE Latin America Transactions*, v. 12, n. 8, p. 1553–1559, 2014.
- NETO, João Estevão Barbosa; DA CUNHA, Jacqueline Veneroso Alves. Masters Committees and Academic Cooperation in Graduate Studies in Accounting. *Contabilidade, Gestão e Governança*, v. 19, n. 1, p. 126-145, 2016.
- NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 2, p. 404–409, 2001.
- NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, v. 101, n. SUPPL. 1, p. 5200–5205, 2004.
- NICHOLAS, J. *Introduction to Descriptive Statistics*. Sydney: Mathematics Learning Centre, University of Sydney, 1999.
- NOORDEN, R. VAN. Global mobility: Science on the move. *Nature*, v. 490, n. 7420, p. 0–3, 2012.
- OECD. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publications: Paris., 2008.
- OECD. Making open science a reality. In: *OECD Science, Technology and Industry Policy Papers*. Paris: OECD Publishing, p. 112, 2015a.
- OECD. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development: The Measurement of Scientific, Technological and Innovation Activities*. Paris: OECD Publishing, 2015b.
- OECD. *OECD Science, Technology and Innovation Outlook 2016*. Paris: OECD Publishing, 2016.
- OKUBO, Y. *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*. STI Working Papers Series 1997/1. OECD, Paris, p.1-70, 1997.
- OMENN, G. S. Grand Challenges and Great Opportunities in Science, Technology, and Public Policy. *Science*, v. 314, n. 5806, p. 1696–1704, 2006.

ORTOLL, E. et al. Principales parámetros para el estudio de la colaboración científica en big science. *Revista Española de Documentación Científica*, v. 37, n. 4, p. 1–11, 2014.

OSTROM, E. *Crossing the Great Divide : Synergy , and Development*. World Development, v. 24, n. 6, p. 1073–1087, 1996.

PACHECO, R. C. S. et al. *Toward CERIF-ScienTI cooperation and interoperability In: Enabling Interaction and Quality: Beyond the Hanseatic League (8th International Conference on Current Research Information Systems)*. Leuven University Press, p. 179, 2006.

PACHECO, R. C. S. *Regimento interno do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina – EGC/UFSC*, 2010.

PACHECO, R. C. S.; SELL, D.; SCHNEIDER, V. *Métodos e Técnicas de Engenharia do Conhecimento: Apresentação da Disciplina. Reflexões da GC sobre EC. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (EGC/UFSC)*, 2013.

PAGE, L. et al. *The PageRank Citation Ranking: Bringing Order to the Web*. p. 1–17, 1998.

PAN, R. K.; FORTUNATO, S. *Author Impact Factor: tracking the dynamics of individual scientific impact*. *Scientific reports*, v. 4, p. 4880, 2014.

PATRICK, W. J.; STANLEY, E. C. *Assessment of research quality*. *Research in Higher Education*, v. 37, n. 1, p. 23–42, 1996.

PELLEG, D. et al. *X-means: Extending K-means with efficient estimation of the number of clusters*. In: *ICML 2000. Proceedings of the 17<sup>th</sup> International Conference on Machine Learning*, v.1, p. 727–734, 2000.

PEREZ-CERVANTES, E.; MENA-CHALCO, J. P.; CESAR JR., R. M. *Towards a quantitative academic internationalization assessment of Brazilian research groups*. In: *E-Science (e-Science). 2012 IEEE 8th International Conference on*. IEEE, p.1-8, 2012.

PERKMANN, M. et al. *Academic engagement and commercialisation: A review of the literature on university–industry relations*. *Research Policy*, v. 42, n. 2, p. 423–442, 2013.

PERLIN, M. S. et al. *The Brazilian scientific output published in journals: A study based on a large CV database*. *Journal of Informetrics*, v. 11, n. 1, p. 18–31, 2017.

- PETERS, M. A. The rise of global science and the emerging political economy of international research collaborations. *European Journal of Education*, v. 41, n. 2, p. 225–244, 2006.
- PLANT, R.; GAMBLE, R. Methodologies for the development of knowledge-based systems, 1982–2002. *The Knowledge Engineering Review*, v. 18, n. 1, p. 47–81, 2003.
- POOCHAOREN, O.; TING, B. Collaboration, Coproduction, Networks – Convergence of Theories. Lee Kuan Yew School of Public Policy Research Paper, v. No. 14-13, n. 65, p. 37, 2014.
- PRESS, W. H. et al. *Numerical Recipes in C*. Cambridge, MA: Cambridge University Press, 1988.
- PRICE, D. J.; BEAVER, D. D. Collaboration in an invisible college. *The American psychologist*, v. 21, n. 11, p. 1011–1018, 1966.
- PRICE, D. J. DE S. *Little science, big science*. New York: Columbia University Press, 1963.
- PRICE, D. S. Networks of scientific papers. *Science*. p. 510–515, 1965.
- PRIEM, J. *Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web*. *First Monday*, v. 15, n. 7, p. 1–14, 2010.
- PRIEM, J. et al. *Altmetrics : A manifesto*. 2010.
- PRIEM, J.; GROTH, P.; TARABORELLI, D. The Altmetrics Collection. *PLoS ONE*, v. 7, n. 11, 2012.
- PRITCHARD, A. Statistical bibliography or bibliometrics? *Journal of Documentation*, v.25, n4, p. 348–349, 1969.
- QUINLAN, J. R. *Induction of Decision Trees*. *Machine Learning*, v. 1, p. 81–106, 1986.
- QUINLAN, J. R. *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1992.
- RAUTENBERG, S. *Modelo de conhecimento para mapeamento de Instrumentos da gestão do conhecimento e de agentes computacionais da engenharia do conhecimento baseado em ontologias*. PhD Thesis, 2009.
- REALPE, A.; WALLACE, L. M. *What is co-production*. London: The Health Foundation, p.1-11, 2010.
- REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-Validation. In: *Encyclopedia of Database Systems*. [s.l.] Springer US, 2009. p. 532–538.

RIBEIRO-JÚNIOR, D. I. Modelo de sistema baseado em conhecimento para apoiar processos de tomada de decisão em ciência e tecnologia. PhD Thesis, 2010.

RICHTER, M. M.; WEBER, R. O. Case-Based Reasoning: A Textbook. Berlin:Springer-Verlag, 2013.

RIESBECK, C. K.; SCHANK, R. C. Inside case-based reasoning. Psychology Press, 2013.

RIP, A. A cognitive approach to science policy. *Research Policy*, v. 10, n. 4, p. 294–311, 1981.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and Empirical Analysis of Relief and RRelief. *Machine Learning*, v. 53, n. 1–2, p. 23–69, 2003.

ROEMER, R. C.; BORCHARDT, R. From bibliometrics to altmetrics: A changing scholarly landscape. *College & Research Library News*, v. 39, n. November, p. 8–9, 2012.

ROSCELLE, J.; TEASLEY, S. The Construction of Shared Knowledge in Collaborative Problem Solving. *Computer-Supported Collaborative Learning*, p. 69–97, 1995.

ROYAL SOCIETY. Machine learning: The Power and promise of computers that learn by example. The Royal Society, p.1-128, 2017.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. Artificial Intelligence. Prentice-Hall Series Englewood Cliffs, 3<sup>rd</sup> ed., 1995.

SALM JUNIOR, J. F. Padrão de projeto de ontologias para inclusão de referências do novo serviço público em plataformas de governo aberto. PhD Thesis, 2012.

SANDSTRÖM, U. Combining curriculum vitae and bibliometric analysis: mobility, gender and research performance. *Research Evaluation*, v. 18, n. 2, p. 135–142, 2009.

SANTOS, P. S. M.; TRAVASSOS, G. H. Scientific Knowledge Engineering: a conceptual delineation and overview of the state of the art. *The Knowledge Engineering Review*, v. 31, n. 2, p. 167–199, 2016.

SARTORI, R. Governança em Agentes de Fomento dos Sistemas Regionais de CT&I. PhD Thesis, 2011.

SCELLATO, G.; FRANZONI, C.; STEPHAN, P. Migrant scientists and international networks. *Research Policy*, v. 44, n. 1, p. 108–120, 2015.

- SCHREIBER, G. et al. Knowledge Engineering and management, The commonKADS methodology. Cambridge, Massachusetts: The MIT Press, 2000.
- SHRUM, W.; GENUTH, J.; CHOMPALOV, I. Structures of scientific collaboration. MIT Press, 2007.
- SINGH, B. K.; VERMA, K.; THOKE, A. S. Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification. *International Journal of Computer Applications*, v. 116, n. 19, p. 11–15, 2015.
- SMALL, H. Co-citation in the Scientific literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science and Technology*, v. 24, n. 4, p. 28–31, 1973.
- SMALL, H. Visualizing Science by Citation Mapping. *Journal of the American Society for Information Science and Technology*, v. 50, n. 9, p. 799–813, 1999.
- SMITH, D. R. Impact factors, scientometrics and the history of citation-based research. *Scientometrics*, v. 92, n. 2, p. 419–427, 2012.
- SNYDER, L. J. The philosophical breakfast club and the invention of the scientist. Dibner Library Lecture, 2011.
- SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, v. 45, n. 4, p. 427–437, 2009.
- SOUZA, I. M. Gestão das Universidades Federais brasileiras: uma abordagem fundamentada na Gestão do Conhecimento. PhD Thesis, 2009.
- SPIER, R. The history of the peer-review process. *Trends in biotechnology*, v. 20, n. 8, p. 357–358, 2002.
- STANFORD UNIVERSITY. Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, 2016.
- STEEL, B. et al. The role of scientists in the environmental policy process: a case study from the American west. *Environmental Science & Policy*, v. 7, n. 1, p. 1–13, 2004.
- STEEL, B. S.; LACH, D.; WARNER, R. Science and Scientists in the U.S. Environmental Policy Process. *The International Journal of Science in Society*, v. 1, n. 2, p. 171–188, 2009.

STOKOLS, D. . et al. The Science of Team Science. Overview of the Field and Introduction to the Supplement. *American Journal of Preventive Medicine*, v. 35, n. 2 SUPPL., p. S77–S89, 2008.

STUART, D. Metrics for an increasingly complicated information ecosystem. *Online Information Review*, v. 39, n. 6, p. 848–854, 2015.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, v. 25, n. 1–2, p. 161–197, 1998.

SUBRAMANYAM, K. Bibliometric studies of research collaboration: A review. *Journal of Information Science*, v. 6, n. 1, p. 33–38, 1983.

SURESH, S. Global challenges need global solutions. *nature*, v. 490, n. 720, p. 337–338, 2012.

TAXWEILER, R.N.N. Um Modelo Para a Extração de Perfil de Especialista Aplicado às Ferramentas de Expertise Location e Apoio à Gestão do Conhecimento. *Knowledge Engineering*. Master Thesis, 2016.

THELWALL, M.; SUD, P. No citation advantage for monograph-based collaborations? *Journal of Informetrics*, v. 8, n. 1, p. 276–283, 2014.

THOMSON-REUTERS. Whitepaper Using Bibliometrics: Thomson Reuters, p. 12, 2008.

TRANFIELD, D.; DENYER, D.; SMART, P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, v. 14, p. 207–222, 2003.

TSAI, C.-T.; LIAO, W.-F. The formation and performance of university technological collaboration: A case of national science and technology program for telecommunication in Taiwan. In: *Management of Engineering & Technology*, 2009. PICMET 2009. Portland International Conference on. IEEE, p. 305-311, 2009.

TUESTA, E. F. et al. Analysis of an Advisor–Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil. *PLoS ONE*, v. 10, n. 5, p. 1–18, 2015.

UNGER, D. D.; RUMRILL JR., P. D. An Assessment of Publication Productivity in Career Development and Transition for Exceptional Individuals: 1978-2012. *Career Development for Exceptional Individuals*, v. 36, p. 25–30, 2013.

VALENDUC, G. et al. Changing careers and trajectories. How individuals cope with organisational change and restructuring, p.1-76, 2009.

- VAN NOORDEN, R. A profusion of measures. *Nature*, v. 465, n. June, p. 864–866, 2010.
- VAN NOORDEN, R. et al. Who is the best scientist of them all? *Nature*, p. 4–8, 2013.
- VARELA, F. G.; MATURANA, H. R.; URIBE, R. Autopoiesis: the organization of living systems, its characterization and a model. *Currents in modern biology*, v. 5, n. 4, p. 187–196, 1974.
- VASILEIADOU, E. Research teams as complex systems and implications for reseach governance. In: *Science and Innovation Policy, 2011 Atlanta Conference on. IEEE*, p.1-9, 2011.
- VINKLER, P. Research contribution , authorship and team. *Scientometrics*, v. 26, n. 1, p. 213–230, 1993.
- WALTMAN, L. A review of the literature on citation impact indicators. *Journal of Informetrics*, v. 10, p. 365–391, 2016.
- WALTON, D.; ZHANG, N. The epistemology of scientific evidence. *Artificial Intelligence and Law*, v. 21, n. 2, p. 173–219, 2013.
- WASSERMAN, S.; FAUST, K. *Social Network Analysis, Methods and Applications*. Cambridge University Press, New York, NY. v.8, 1994.
- WATSON, I. Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, v. 12, n. 5–6, p. 303–308, 1999.
- WEBER-LEE, R. et al. Using typicality theory to select the best match. In: *European Workshop on Advances in Case-Based Reasoning*. Springer, Berlin, Heidelberg, v. 1168, p. 445–459, 1996.
- WELSH, E.; JIROTKA, M.; GAVAGHAN, D. Post-genomic science: Cross-disciplinary and large-scale collaborative research and its organizational and technological challenges for the scientific research process. *Philosophical Transactions of the Royal Society A*, v. 364, n. 1843, p. 1533–1549, 2006.
- WETTSCHERECK, D.; AHA, D. W.; MOHRI, T. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review*, v. 11, p. 273–314, 1997.
- WHITE, H. D.; GRIFFITH, B. C. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, v. 32, n. 3, p. 163–171, 1981.
- WIELINGA, B. J.; SCHREIBER, A. T.; BREUKER, J. A. KADS: a modelling approach to knowledge engineering. *Knowl. Acquis.*, v. 4, n. 1, p. 5–53, 1992.

WIENER, N. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, 1961.

WILDGAARD, L. E.; et al. Can we implement the Leiden Manifesto principles in our daily work with research indicators? Report from the 5<sup>th</sup> meeting of the Danish Research Indicator Network (FIN), at Copenhagen University Library - Frederiksberg. 2016

WILSDON, J. et al. *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. Sage, 2016.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2<sup>nd</sup> ed., Morgan Kaufmann Publishers, 2005.

WOELERT, P.; MILLAR, V. The “paradox of interdisciplinarity” in Australian research governance. *Higher Education*, v. 66, n. 6, p. 755–767, 2013.

WOOLLEY, R.; CAÑIBANO, C.; TESCH, J. *A Functional Review of Literature on Research Careers: Ingenio Working Paper Series (CSIC-UPV)*. 2016

WRAY, K. B. Did professionalization afford better opportunities for young scientists? *Scientometrics*, v. 81, n. 3, p. 757–764, 2009.

XUE, M. X. M.; ZHU, C. Z. C. A Study and Application on Machine Learning of Artificial Intelligence. 2009 International Joint Conference on Artificial Intelligence, p. 272–274, 2009.

YAGI, E.; BADASH, L.; DE BEAVER, D. B. Derek J. de S. Price (1922–83): Historian of science and herald of scientometrics. *Interdisciplinary Science Reviews*, v. 21, n. 1, p. 64–76, 1996.

YOUTIE, J.; BOZEMAN, B. Social dynamics of research collaboration: norms, practices, and ethical issues in determining co-authorship rights. *Scientometrics*, v. 101, p. 953–962, 2014.

YU, L.; LIU, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, v. 5, n. October, p. 1205–1224, 2004.

ZHANG, C. . et al. Research collaboration in health management research communities. *BMC Medical Informatics and Decision Making*, v. 13, n. 1, 2013.

## APPENDIX A

Frame 31 – The rank of 100 most similar *fit candidates* for the purpose of the Experiment II – Part II

#	Fit candidate	PhD Year	Years since PhD	Institutional Affiliation	Field of Affiliation	Similarity score
1	7550	2003	14	Southeast, Midwest	Biological Sciences	0.9667
2	9782	1998	19	Southeast, Midwest	Biological Sciences	0.9662
3	6004	1996	21	Southeast, Midwest	Biological Sciences	0.9658
4	3987	1997	20	Southeast, Midwest, North	Biological Sciences	0.9657
5	10736	2000	17	Southeast, Northeast	Biological Sciences, Health Sciences	0.9656
6	333	2002	15	Southeast	Health Sciences	0.9656
7	8257	2000	17	Northeast	Health Sciences	0.9655
8	8438	1996	21	Northeast	Health Sciences	0.9655
9	10556	2000	17	Southeast, Midwest	Biological Sciences, Health Sciences	0.9654
10	5335	1999	18	Southeast	Health Sciences	0.9654
11	4889	1997	20	Southeast	Health Sciences	0.9654
12	7612	1992	25	Southeast	Biological Sciences, Health Sciences	0.9654
13	4678	1994	23	South, Midwest, Northeast	Health Sciences	0.9653
14	5558	1996	21	Southeast, Midwest	Biological Sciences	0.9653
15	3575	1991	26	Southeast	Health Sciences	0.9651
16	3578	1998	19	Southeast, Midwest	Health Sciences	0.9651
17	9301	1999	18	Northeast	Health Sciences	0.9651
18	15024	2002	15	Southeast	Health Sciences	0.9651
19	5224	1995	22	Northeast	Biological Sciences, Health Sciences	0.9650
20	8111	1996	21	South	Health Sciences	0.9650
21	5810	2000	17	Midwest	Health Sciences	0.9649

(cont.)

#	Fit candidate	PhD Year	Years since PhD	Institutional Affiliation	Field of Affiliation	Similarity score
22	3862	1996	21	Southeast	Biological Sciences	0.9648
23	5568	2000	17	South, Midwest	Biological Sciences	0.9648
24	1610	1992	25	Southeast	Health Sciences	0.9647
25	10408	2003	14	Southeast, Midwest	Health Sciences	0.9647
26	10893	1995	22	Northeast	Health Sciences	0.9647
27	2939	2003	14	Midwest	Health Sciences	0.9646
28	4539	2002	15	South, Midwest	Health Sciences	0.9646
29	7566	1987	30	Southeast	Health Sciences	0.9646
30	7132	2000	17	Northeast	Health Sciences	0.9645
31	8534	1999	18	Southeast, Midwest	Health Sciences	0.9645
32	9603	1999	18	Southeast	Health Sciences	0.9645
33	11040	1998	19	Southeast, Midwest	Biological Sciences	0.9645
34	15065	1999	18	Midwest, Northeast	Health Sciences	0.9645
35	3968	2001	16	Southeast	Health Sciences	0.9644
36	9137	2001	16	Southeast, Midwest	Health Sciences	0.9644
37	9178	1996	21	Southeast	Health Sciences	0.9644
38	11274	1994	23	Southeast, Midwest, Northeast	Health Sciences	0.9644
39	10309	1980	37	Southeast	Biological Sciences	0.9643
40	11770	1997	20	Southeast	Health Sciences	0.9643
41	13656	1999	18	Southeast	Biological Sciences	0.9643
42	14365	1997	20	Northeast	Health Sciences	0.9643
43	1294	2001	16	Southeast	Health Sciences	0.9642
44	2948	1992	25	Southeast	Biological Sciences, Health Sciences	0.9642
45	3583	1999	18	Southeast	Health Sciences	0.9642

(cont.)

#	Fit candidate	PhD Year	Years since PhD	Institutional Affiliation	Field of Affiliation	Similarity score
46	4526	1994	23	Southeast, Midwest, Northeast, North	Health Sciences	0.9642
47	4624	2005	12	Northeast	Health Sciences	0.9642
48	5040	1997	20	Southeast	Health Sciences	0.9642
49	2081	2000	17	Midwest	Health Sciences	0.9641
50	2081	2000	17	Southeast	Health Sciences	0.9641
51	2365	2002	15	South, Southeast, Midwest, Northeast, North	Health Sciences	0.9641
52	5590	2006	11	South, Southeast, Midwest	Health Sciences	0.9641
53	9347	2005	12	Midwest	Health Sciences	0.9641
54	12142	1992	25	Southeast, Midwest	Biological Sciences	0.9641
55	1631	1993	24	Southeast	Health Sciences	0.9640
56	2191	1991	26	Southeast	Health Sciences	0.9640
57	3223	1989	28	Southeast	Health Sciences	0.9640
58	3341	1993	24	South, Northeast	Biological Sciences	0.9640
59	10480	2001	16	Southeast	Health Sciences	0.9640
60	11478	1987	30	Southeast, Midwest	Biological Sciences	0.9640
61	12203	1994	23	South, Southeast, Midwest, Northeast	Biological Sciences	0.9640
62	12863	2000	17	South	Health Sciences	0.9640
63	13830	2004	13	South	Health Sciences	0.9640
64	14010	1992	25	Southeast	Biological Sciences	0.9640
65	14108	2002	15	Midwest	Biological Sciences, Health Sciences	0.9640
66	1346	2003	14	Southeast	Health Sciences	0.9639
67	2727	1997	20	South, Southeast, Midwest	Health Sciences	0.9639

(cont.)

#	Fit candidate	PhD Year	Years since PhD	Institutional Affiliation	Field of Affiliation	Similarity score
68	2731	1989	28	Southeast, Midwest	Biological Sciences	0.9639
69	10763	1999	18	Southeast	Biological Sciences	0.9639
70	15064	1997	20	Southeast	Health Sciences	0.9639
71	2961	2001	16	Southeast	Health Sciences	0.9638
72	5795	1989	28	Southeast	Health Sciences	0.9638
73	8567	2000	17	Southeast	Health Sciences	0.9638
74	12057	1981	36	Southeast	Health Sciences	0.9638
75	14284	1994	23	Southeast, Midwest	Biological Sciences	0.9638
76	1169	1996	21	Northeast	Health Sciences	0.9637
77	1555	1995	22	Southeast, Midwest	Health Sciences	0.9637
78	2937	1993	24	Northeast	Biological Sciences	0.9637
79	3277	2000	17	Southeast	Health Sciences	0.9637
80	3877	1978	39	Southeast	Health Sciences	0.9637
81	4515	1990	27	Southeast	Health Sciences	0.9637
82	10379	1997	20	Northeast	Biological Sciences	0.9637
83	10993	2001	16	Southeast, Midwest, Northeast	Biological Sciences	0.9637
84	11160	2004	13	Southeast	Health Sciences	0.9637
85	11530	1997	20	South, Southeast, Midwest	Health Sciences	0.9637
86	3017	2001	16	Southeast, Midwest	Biological Sciences	0.9636
87	3989	2001	16	Southeast, Midwest	Health Sciences	0.9636
88	6390	1995	22	Northeast	Health Sciences	0.9636
89	7972	2003	14	Southeast	Health Sciences	0.9636
90	11285	1995	22	Southeast	Health Sciences	0.9636
91	13275	1997	20	Southeast, Midwest	Health Sciences	0.9636
92	14513	2001	16	Southeast	Health Sciences	0.9636

(cont.)

#	Fit candidate	PhD Year	Years since PhD	Institutional Affiliation	Field of Affiliation	Similarity score
93	2539	2000	17	Southeast	Health Sciences	0.9635
94	3497	2000	17	Midwest, Northeast	Health Sciences	0.9635
95	7595	2002	15	South, Southeast, Midwest, Northeast	Health Sciences	0.9635
96	8823	2003	14	Southeast, Midwest	Biological Sciences, Health Sciences	0.9635
97	10210	1997	20	South	Health Sciences	0.9635
98	14936	1994	23	Southeast	Health Sciences	0.9635
99	893	1991	26	Southeast	Biological Sciences	0.9634
100	1463	1996	21	North	Biological Sciences	0.9634

Source: The author, 2017.