Goal-Oriented Requirements Engineering: A Systematic Literature Map

Jennifer Horkoff*, Fatma Başak Aydemir†, Evellin Cardoso†, Tong Li†, Alejandro Maté†‡, Elda Paja†, Mattia Salnitri†, John Mylopoulos†, Paolo Giorgini†
*City University London, UK, horkoff@city.ac.uk
†University of Trento, Italy
{aydemir,tong.li,paja,salnitri,jm,paolo.giorgini}@disi.unitn.it, evellin.souzacardoso@unitn.it
‡University of Alicante, amate@dlsi.ua.es

***Abstract*—Over the last two decades, much attention has been paid to the area of Goal-Oriented Requirements Engineering (GORE), where goals are used as a useful conceptualization to elicit, model and analyze requirements, capturing alternatives and conflicts. Goal modeling has been adapted and applied to many sub-topics within RE and beyond, such as agent-orientation, aspect-orientation, business intelligence, model-driven development, security, and so on. Despite extensive efforts in this field, the RE community lacks a recent, general systematic literature review of the area. As a first step towards providing a GORE overview, we present a Systematic Literature Map, focusing on GORE-related publications at a high-level, categorizing and analyzing paper information in order to answer several research questions, while omitting a detailed analysis of individual paper quality. Our Literature Map covers the 246 top-cited GORE-related conference and journal papers, according to Scopus, classifying them into a number of descriptive paper types and topics, providing an analysis of the data, which is made publicly available. We use our analysis results to make recommendations concerning future GORE research.**

Keywords*—*requirements engineering; goal model; systematic literature survey; systematic literature map; evidence-based requirements engineering

I. Introduction

The quality of a software system critically depends on the degree to which it fulfills its requirements. Such requirements are often elicited, modeled and analyzed as (stakeholder) goals. The field of Goal-Oriented Requirements Engineering (GORE) has emerged, particularly in the last two decades. Typically, goals are elicited and conceptualized in terms of some form of model. Goal models have been used as an effective means for capturing the interactions and tradeoffs between requirements, but they have also been used more broadly: in Software Engineering, Information Systems, Conceptual Modeling, and Enterprise Modeling.

In this work, we aim to understand the landscape and status of existing work in GORE at a high-level. In a recent RE meta-survey, Bano et al. have pointed out that there has yet to be a systematic literature review of GORE publications [3]. Although a few GORE reviews exist, e.g., [13], [1], they are focused on sub-topics or frameworks within GORE, and not the area in its entirety.

We focus our investigation on a set of particular research questions. Broadly speaking, we are interested in mapping the space of GORE research. More specifically, we are interested in classifying the types of GORE publications (proposals, extensions, meta-studies, etc.), the nature of research evaluation, common topics appearing in GORE work, common frameworks, publication venues, and citation distributions. Overall, we ask if interest in GORE is increasing or decreasing? We make an initial attempt to understand the reasons behind the results of our mapping.

Kitchenham et al. have advocated Evidence-based Software Engineering [17], inspired by Evidence-based Medicine. Their work finds and assesses available evidence to address software engineering questions for researchers and practitioners in a systematic fashion. In our study, we perform Evidence-based Requirements Engineering (EBRE), systematically finding and summarizing available publications in order to answer goal model-related research questions.

Specifically, we produce a Systematic Literature Map (SLM) summarizing publications falling under the scope of our study without considering their quality [18], [23]. This SLM can be beneficial for several types of readers. For researchers interested in GORE, the map helps to build upon existing work, avoiding the proverbial 'reinvention of the wheel', helping to understand trends, and guiding efforts in new directions. For practitioners, this map offers ideas on the most prominent GORE methods and frameworks, including pointers to work containing further details.

This work follows the same theme as previous work by some of our authors presented in [12], [13], but with a different focus and method. These papers provided a SLM [12] and then a Systematic Literature Review (SLR) [13] focusing specifically on methods which transform or map to or from goal-oriented methods, a subset of the focus of this paper. These previous surveys found papers through a mix of systematic search and reference "snowballing", without using the number of citations as an excluding criterion. In this work we use only systematic search, with a citation cutoff to manage survey size. Despite the broader scope of the current survey, because of these differing methods of finding papers, the publications included in the current survey are not a super-set of the papers

in the previous surveys, i.e., most publications included in these previous surveys are not included in the current SLM. Specifically, the overlap between the 170 papers included in [12] and the 246 papers included in this survey is 29 papers, while the overlap with the 247 papers in [13] is 40.

The rest of the paper is organized as follows. We first introduce our research questions in Sec. II, then describe the scope, classification schema and key terminology of our study in Sec. III. Sec. IV presents the methodology followed. Sec. V summarizes the results of the SLM, while Sec. VI discusses survey results and design alternatives. Sec. VII lists threats to the validity of the study. Sec. VIII reviews related work while, Sec. IX offers conclusions and ideas for future work.

II. Research Questions

As per Petersen et al. [23], we articulate the specific research questions (RQs) guiding our SLM. Our overarching aim is to map the landscape of highly-cited GORE research. We ask more detailed questions, as listed in Table I. It is important to note that our unit of analysis is publications, and not research approaches (e.g., frameworks such as KAOS, i*, Secure Tropos). Focus on approaches would be interesting, but is subject to much interpretation, see Sec. VI for a discussion of alternative survey approaches.

III. Scope and Preliminaries

In this section, we provide definitions of key concepts used to define the scope and classification schemes of our SLM.

A. Key Terms

We define *Goal-Oriented Requirements Engineering* as the study or application of goal models in Requirements Engineering. A *goal model* is a model expressed in a goal-oriented language. Such languages include the concept of goal as a first class object, are often graphical and come with a visual syntax (e.g. i* [32], KAOS [7]) but may also be textual (e.g., GBRAM [2]). We adopt the notion of a *language* from [10]: "a language consists of a syntactic notation (syntax), which is a possibly infinite set of legal elements, together with the meaning of those elements, which is expressed by relating the syntax to a semantic domain." Languages can be graphical or textual, and the semantics (meaning) can be formally or informally defined.

Although we focus on the use of goal models in Requirements Engineering, we do not exclude those publications which are either aimed for different research fields, or which apply goal models to a new context, as long as the authors relate their work back to GORE. See the description of our systematic search string in Sec. IV for more information.

B. Inclusion and Exclusion Criteria

We focus our investigation on publications appearing in international journals, conferences, or symposia. We omit theses, focusing on work which has been published in international venues. Among venues, we exclude workshop publications and regional conferences. Our scoping criteria are summarized in Table II.

C. Classification Schemes

We have collected basic information for each included publication. Some of this information was extracted automatically from Scopus, while the rest was added (or corrected) by hand. For each included publication, we kept track of: the paper title, authors with their affiliations and countries, venue, type of venue, year, number of citations (according to Scopus, Google Scholar, and Web of Science), number of pages, and GORE framework (e.g., i*, KAOS). In cases when the GORE framework was not clear or multiple frameworks were applied, we used the tags "general" or "multiple".

In addition to this basic information, we have endeavored to understand GORE publications via two classification schemes. The first, the *type* of paper refers to the research contributions, methods and/or structure provided by the paper. Our schema bears similarities to the classifications scheme of Wieringa et al. [31]; however, after our experiences using this scheme in [12], [13], and based on our discussions while classifying papers, we have designed a slightly broader, more descriptive scheme. The second classification refers to the *topic* of the paper (e.g., Scenarios, Agile, NFRs) , independent from the research method. Here, we used a bottom-up approach, performing an initial assessment of a subset of GORE-related papers, deriving a set of commonly covered topics.

We call the process of applying these categories to papers "tagging" or "coding", as per the typical terminology of qualitative coding or tagging, applying one or more "tags" or "codes". When tagging a paper, we tried to be true to the terminology used by the authors, e.g., if the authors say they extended goal models with scenarios, we would include the *extension* tag as a paper type and the *scenario* tag as a paper topic. The selection, classification and inter-coder agreement on these schemes is discussed more in Sec. VI.

For space considerations, we provide the detailed definitions of our types and topics online[1], but provide a list of all tags, with associated keywords for paper topics, in Tables III and IV. We highlight particular tags whose interpretation may be less obvious. For instance, for us a proposal was any publication that makes a new contribution, e.g., a language, extension, integration, or algorithm. We tagged a paper as a formalization if it contained axioms or some formal logical language relating to the proposal, particularly looking for logical operators (e.g., $\neg$, $\vee$, $\Rightarrow$). Meta studies were papers such as this one, which provided a significant overview or study of existing work. If the publication mentioned a tool or implementation which facilitated the contribution, we tagged with implementation. We had several evaluation categories, most of which were self-evident, with the exception of Case Study. We used this tag when the publication included a case study which evaluated a claimed contribution. Whether the reported case study was a case study or only an illustrative example depended on depth and realness: if the case was detailed, extracted from a real world problem, or if there was more detailed information

[1] http://www.cs.toronto.edu/goreslm/PublicationTypesTopics.pdf

TABLE I: Research Questions

RQ1	How can we classify the type of GORE approach? Can GORE publications be classified as proposals, formalizations, meta-studies, integrations, extensions, ontological interpretations? Do they contain an implementation? How has this changed over time?
RQ2	Do GORE publications contain evaluation? Of what type? How has this evolved?
RQ3	What are the topics covered by GORE publications? How have these topics evolved over time?
RQ4	What goal modeling frameworks have been used in the publications?
RQ5	In what journals or conferences do approaches typically appear?
RQ6	What techniques are most widely cited? Are citations equally distributed? How do they vary per citation source?
RQ7	Is interest in GORE increasing or decreasing?

TABLE II: Publication inclusion and exclusion criteria

Inclusion Criteria	**Exclusion Criteria**
Has a significant component that deals with GORE	Does not significantly relate to GORE or
In conference, journal, or in/is a book, and	Is a thesis, workshop or regional conference, or
Is published in English, and	Is published in another language, or
Is more than 3 pages.	Is 3 pages or less.

TABLE III: Publication Types

Paper Type
Proposal
Formalization
Meta Study
Implementation
Integration/Transformation/Mapping
Extension
(Ontological) Interpretation
Evaluation (Benchmark)
Evaluation (Controlled Experiment)
Evaluation (Questionnaire)
Evaluation (Case Study)
Evaluation (Scalability)

TABLE IV: Publication Topics and the Corresponding Keywords

Paper Topic	**Topic Keywords**
Agent	agent, actor
Aspects	aspect
Business Intelligence/Modeling	business intelligence, business modeling, KPI, indicator, enterprise modeling, strategic management
Conflicts	conflict
Requirements Engineering	requirements engineering, RE, requirements
Early Requirements Engineering	early, early RE, early requirements engineering
Model driven Development	model–driven, MDD, MDE, MDA
Non-Functional Requirements/Softgoals	softgoal, NFR, non-functional
Privacy, Security, Risk & Trust	privacy, security, risk, trust
Systematic Reasoning	reasoning, analysis, automated, propagation, evaluation, metrics
Adaptation & Variability	adapt, adaptation, adaptability, variability, evolution, autonomic
Architecture	architecture
Compliance	compliance, law, policy, regulations
Patterns	pattern
Agile	agile, scrum, lean, extreme, XP
Scenario	scenario sequence, use case

available in another source, typically we classified the paper as containing a case study.

In terms of topics, in most cases, if the paper involved significantly the particular topic we selected that topic. Significance involved a degree of judgment, but was an aspect we discussed over several iterations. The Business Intelligence tag was a somewhat all-encompassing tag covering both business intelligence, analysis, and business modeling, as we had trouble clearly differentiating between the possible sub-categories. Systematic reasoning also required some care. Here we selected publications which contained algorithmic or mathematical analysis of a model to answer some question or find some property. This could be formal, qualitative, quantitative, automated, interactive, or manual, as long as it is systematic and repeatable.

IV. Survey Method

In this section we describe the pre-survey preparation, survey steps, and post-survey processing.

A. Pre-Survey Preparation

Database support. As our survey was designed as a SLM, we anticipated we would have the need to store and process a significant quantity of data. The initial steps of our planning involved designing an extendable and adjustable database schema for the publication reviews. The database technology was built with MySQL, with a front-end in HTML and PHP, allowing us to view all papers and add information for particular papers[2].

Initial Tags. We started with an initial conceptualization for the paper type scheme (Sec. III-C), but with the paper topics, we performed a grounded analysis, inspired by grounded theory [27]. We started with a set of papers we knew to be related to GORE and then "snowballed" through the papers, following the reference links to other related papers, assigning type and topic tags to each paper, and proposing our own perceived topics. The tagging processes ended when we got to a set of 110 papers, adopting the set of topics generated so far. In a group discussion, we processed the topics, merging similar ones, also coming up with definitions, resulting in an early version of the list in Table IV. This process also helped us to refine the paper type scheme, with unclear types removed or refined.

Publication Processing. It was necessary to work out clear guidelines as to how and to what degree to read selected publications. Given the high-level, mapping nature of our survey, it was not necessary to carefully read in its entirety

[2]See http://www.cs.toronto.edu/goreslm/DBInterface.png for a screenshot of our SLM database interface

each paper in the survey. Many literature maps restrict reading to the abstract or introduction. We decided to read the title, abstract, introduction and conclusion. The reader/tagger could optionally flip through the details of the paper, particularly section headings, to make clarifications or resolve questions. As most papers were about modeling, perusing was particularly useful to see the details of the included model(s).

Inter-coder agreement. At this point, it was necessary to evaluate how consistently the coders could apply the type or topic tags. We performed two rounds of inter-coder reliability (ICR) tests. For these rounds we used papers randomly selected from our goal-related thesis bibliographies. As we estimated our final set of papers would be approximately 300, we chose a set of 30 papers, making up about 10% of the final size. We did not select 10% of the exact set we would eventually process as at this point we believed this snowballing set would make up a significant part of our survey. Survey design choices are discussed in Sec. VI.

The initial team of paper taggers was made up of seven post-docs and graduate students with some association to the University of Trento and some experience with goal modeling. In the first round of ICR testing all seven coders coded a set of 30 potential GORE papers. We evaluated ICR on the types and topic tags using Krippendorff's alpha, which indicates our coding consistency per code across all 30 papers [19]. As our codes are overlapping, the commonly applied Kappa measures (Cohen's, Fleiss') are not applicable, also Krippendorf's alpha gives us the benefits of showing specifically which codes we perform well or poorly on. Here we aim for an agreement level minimum of 0.67, ideally greater than 0.80, as per [19].

The mean, median, min and max ICR scores for round one are shown in Table V. Note that these scores do not measure only agreement, but agreement accounting for chance, so a score of zero does not mean we did not agree, but that we are as accurate as choosing values randomly. These scores were obviously not optimal for several tags. We repeated the process in round two for a different set of 25 papers randomly selected from the same sources. Before performing this second round we took several actions: 1) we had extensive discussions on the meanings of tags with scores < 0.8, coming up with shared text definitions for all tags, 2) we dropped and merged some tags which caused confusion, 3) we dropped a coder with background less-related to GORE, and 4) we tried to better emulate the final process by having all codes checked by a second person. For the last point, after each coder had coded each paper, we assigned a second coder to each paper, such that each coder would be a pair with each other coder the same amount of times. The second coder checked the tags of the first, and disagreements were discussed. The results for round two are shown in the second row of Table V.

TABLE V: Krippendorff's Alpha ICR Results

		Paper Types				Paper Topics			
Round	**Pub#**	**Mean**	**Med**	**Min**	**Max**	**Mean**	**Med**	**Min**	**Max**
1	30	0.62	0.66	0.08	1.0	0.42	0.40	0	1.0
2	25	0.74	0.79	0.5	0.88	0.63	0.61	0.19	1.0

Obviously, some of the tags still had less than optimal agreement. We went through a second round of group discussions, refining definitions, changing and adding some further tags. Due to time constraints, we opted not to do yet another round of ICR coding and testing. As this process had already taken six months (see Sec. VI for a discussion of why), we were not convinced that extra time would be worth the possible increase in scores, ICR scores are discussed further in Sec. VII.

B. Systematic Search

In order to reduce potential bias in selecting among candidate publications, we moved our focus from snowballing to systematic search. After evaluating various potential sources, including Google Scholar and Web of Science, we decided to perform our search through Scopus, as it covers major publishers in RE (ACM, Springer, IEEE), is more inclusive than Web of Science, but less inclusive than Google Scholar, which may include many non-peer reviewed papers such as technical reports. Note that although we perform our publication search using only Scopus, we extract and compare citation data from Scopus, Google Scholar and Web of Science.

We derived our search string from our research questions, searching the title, abstract and keywords for : ("goal-oriented" OR "goal model" OR "goal modeling" OR "goal modelling") AND "requirements", limiting the search to conference proceedings, book chapters, (journal) articles, or articles in press. As of 2015-12-16, we found 966 results. The next step was to automatically import as much information from Scopus as possible into our Database. Elements such as the publication name, authors, venue, year, page numbers and affiliations were imported using a script.

It was clear that it was not feasible to evaluate all 966 papers; furthermore, we found that many papers had a very small number of citations according to Scopus (394/966 papers, 41%, had 0 citations). We chose to evaluate all publications having three or more citations according to Scopus, evaluating a total of 350 publications. During our paper processing, we found 104 papers that were out of the scope according to our criteria, ending up with 246 papers for our study.

To evaluate these papers, we adopted the following process: we divided the papers up into six equal groups, sorting by number of citations then assigning every sixth paper to a group. Each group was given to a coder, who processed all papers in her group with ≤ 3 citations. This means that each coder had to process about 60 papers. When the process was complete, each paper was reassigned to a second coder, for a cross-check. We assigned the papers such that every coder was checking a roughly equal number of papers coded by each other coder. The second coder also reviewed papers and tags, raising issues in various fields when they thought a code was missing or incorrect. Issues were stored in the database and were discussed and resolved. Overall, we found and resolved 182 issues concerning 124 out of 246 papers.

Finally, we performed a round of data cleaning to check and resolve missing fields or any remaining issues. The first review stage took our coders about a month, while the second

round took about three weeks. It took each coder anywhere from 10 to 30 minutes to process each publication.

C. Post-Survey Analysis

Although we are focusing our analysis on particular analysis questions, given the large scope of data that has been collected for each publication, future dimensions of analysis should be supported[3]. In order to support current and future data analysis, we made use of OLAP (Online Analytical Processing) analysis provided by Business Intelligence tools [21], [11], as it provides a flexible analysis based on multidimensional modeling that does not require re-processing the data. A multidimensional model required for supporting the analysis has been derived from the RQs by following data warehousing construction methodologies [20]. We used this method to help us find and present the data to address our RQs, presented in the next section.

V. Literature Map Results

We present the data for each of our RQs with an emphasis on visual maps and graphs, as is recommended for SLMs [23].

RQ1. We summarize the number of classifications for our 246 papers in Fig. 1. Our classifications are overlapping, we have 938 paper types over 246 papers, an average of 3.8 type tags per paper. We can see that nearly all (91%) of papers propose something new, while there is a near even number of extensions, formalizations and integrations/transformations/mappings around 40%. About 40% of the publications offer some sort of formalization, and nearly half, 49% offer some sort of implementation. Ontological interpretations are relatively rare (5%), as are Meta Studies (9%). Overall, the focus seems to be on proposing independent new methods, while only making extensive use of past approaches less than half of the time.

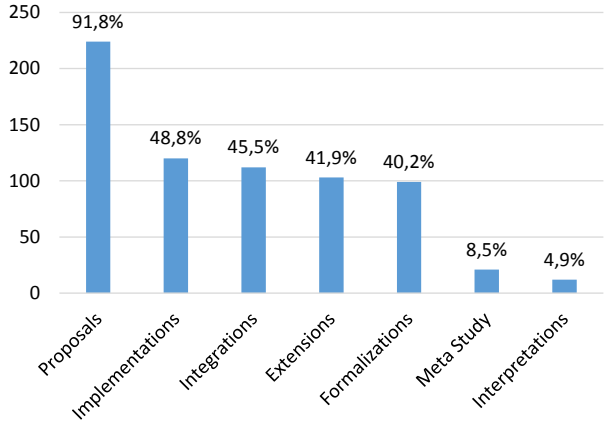

Fig. 1: Count of Paper Types (**RQ1**)

We can gain further insights by tracking this data over time (Fig. 2). The top line shows the number of papers per year, as a comparison. For this and any other graph showing information per year, we must account for the fact that our mapping includes only those publications with more than three citations. Thus we have a bias towards older publications, while newer publications are less likely to be included. This must be accounted for when considering the drop in all data from 2013-15.

We see the proposals hold steady with the total number of papers, while most other types of papers hold at about half the total number of papers. Implementations seem to be on the rise, with the number of integrations and extensions also appearing to rise slightly, all relative to the number of papers. This breakdown gives a slightly more optimistic view, with incorporation of past methods seemingly on the rise.

RQ2. Overall, 53% of the 246 papers contain a case study, as per our tag definition, 27% some evaluation of scalability, 7% a controlled experiment, 7% questionnaires, and 4% contain some type of benchmark. The evolution of these tags over time is shown in Fig. 3. In general, the rise and fall of each type of evaluation follows the pattern of number of papers per year. We can see some low points in the evaluation of Scalability relative to the number of papers, while empirical studies other than Case and Scalability studies are low overall. It appears the use of Case Studies may be on a slight rise, as the slope of the number of papers is steeper than the Case Study slope beyond 2012. For example, in 2008 44% of papers have case studies, compared to 54% in 2012. Future analysis is needed to determine whether this trend continues to hold.

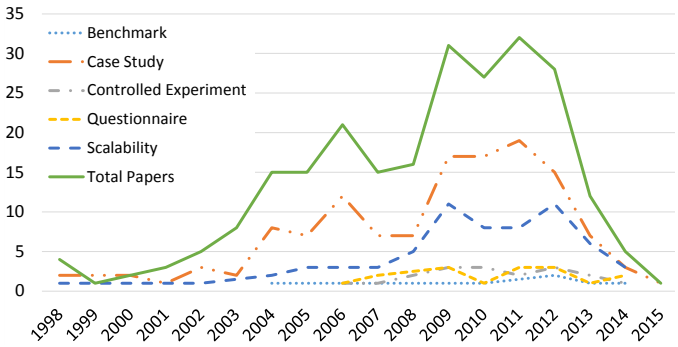

Fig. 3: Evaluation Types Per Year (**RQ2**)

RQ3. Fig. 4 shows the breakdown of paper topics, we have 910 topics over 246 papers, with an average of 3.7 topics per paper. We can see that most papers (91%) involve RE, unsurprisingly given our search string, but 9% of papers were significantly out of the RE field. Other popular topics include Agents (50%), Reasoning (43%), and NFR/Softgoals (36%). We show a breakdown of the top five topics per year in Fig. 5, starting from 2000. Examining trends in these popular topics, the focus seems to rise and fall with the general number of papers, with a few exceptions. Interest in Reasoning seems to have decreased relatively between 2009-11, but seems to have increased relatively in 2012. Interest in Adaptation/Variability/Evolution has increased recently relative to other topics, possibly accounting for the latest spike in overall GORE papers. NFR/Softgoal interest appears to be decreasing.

[3]The raw data for our database is available here: http://www.cs.toronto.edu/goreslm/GOLSurvey-21050304.sql

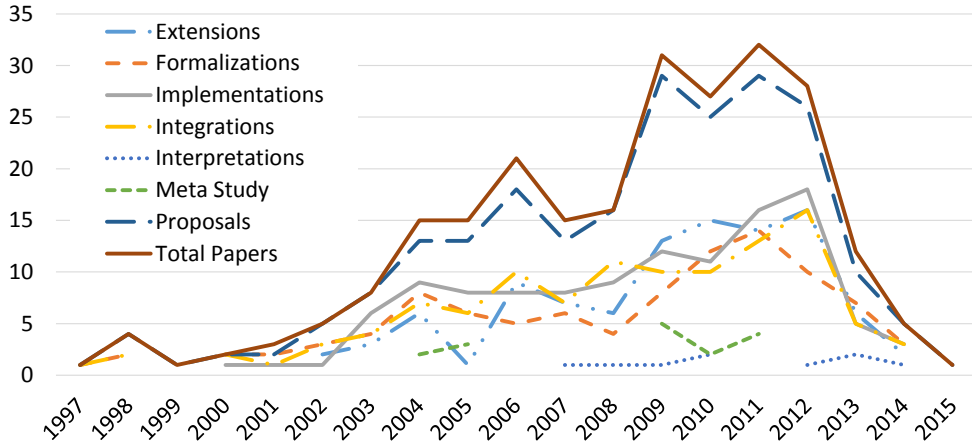

Fig. 2: Paper Types per Year (**RQ1**)

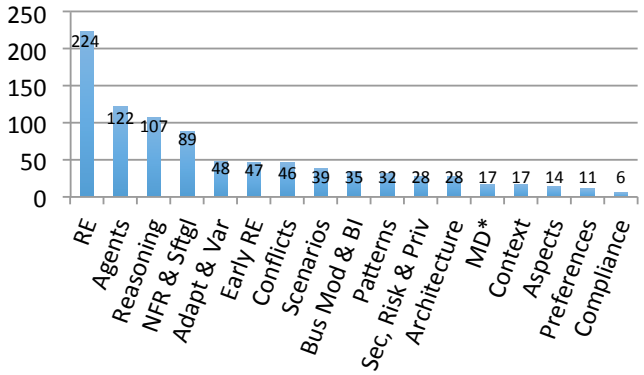

Fig. 4: Total Paper Topics (**RQ3**)

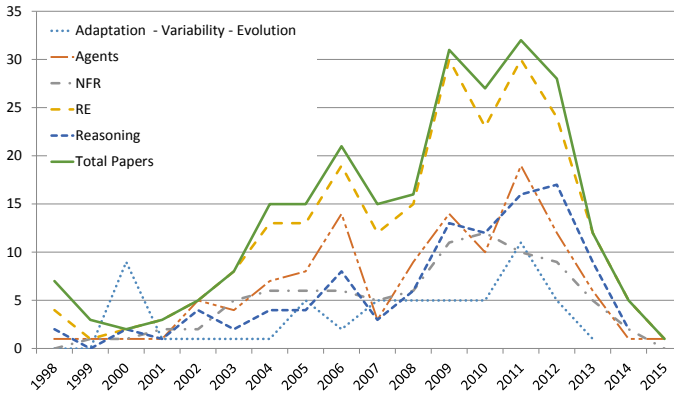

Fig. 5: Top Five Total Paper Topics Per Year (**RQ3**)

RQ4. Fig. 6 shows the GORE frameworks used in our included publications. As the frameworks tag does not overlap, we can view this in a pie chart. We can see that although KAOS and i* appear in near to the same number of publications (13%), the most popular choice is to use goal modeling in general, without committing to a particular framework. It's also fairly common (7%) to significantly use multiple frameworks in one paper.

RQ5. Our SLM found a total of 111 unique venues. We show the top 12 publication venues in Fig. 7, each with five or more publications. We can see that the RE conference dominates, followed by REJ, then other conferences and journals with roughly equal paper numbers. 107 out of 246 (43%) publications in our SLM appear in one of these top 12 venues, meaning that the spread of publication venues is still quite wide. This can make it difficult to consolidate and share GORE-knowledge, but also helps to demonstrate the uses of GORE beyond the RE community.

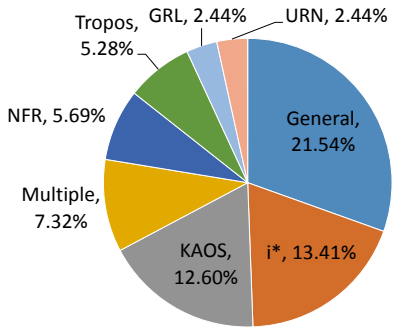

Fig. 6: Frameworks Used in 246 Publications (**RQ4**)

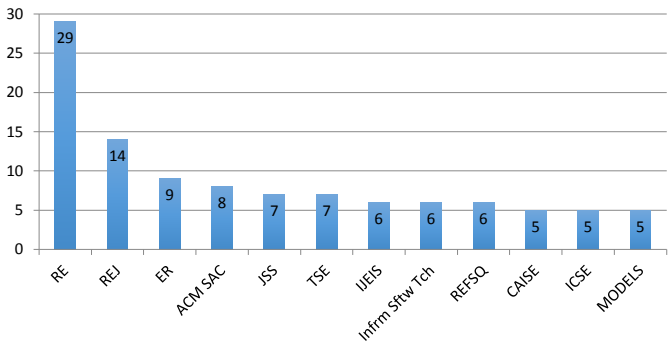

Fig. 7: Top 12 Publication Venues (**RQ5**)

RQ6. We show the top 20 cited papers as per Scopus in Fig. 8[4]. We see that the Google Scholar citations for van

[4]Full paper references can be found here: url-http://www.cs.toronto.edu/goreslm/Papers.html

Lamsweerde's 2001 Guided Tour [28] dominates all other citations. Although this paper is also the most cited in Scopus, the differences between it and others are not as large, highlighting the different algorithm that Google uses to count citations. We show an alternative, more readable version omitting Google Scholar results in Fig. 9. Here we can see that there are a few highly-cited papers, while citations for the other papers tail off gradually. We see this as a common phenomenon in a research area, where a few papers become seminal and are the default, "go-to" citation for an area. As mentioned in Sec. IV, 41% of the 966 papers had zero citations, and 616 out of 966 (64%) papers had less than three citations. Of these 616 papers, only 242 are recent, from 2013-15. This means there are many older GORE-related papers which are not highly cited.

In general, these charts highlight the differences between citation sources. If possible, it is best to consider multiple sources of citations when analyzing publication data. In our case, we have collected all three data points, but focus on Scopus as a data source which is intermediate when compared to Google Scholar or Web of Science.

RQ7. When looking at overall interest in GORE, we can refer to Figs. 2, 3 and 5, each of which shows the total number of GORE papers per year included in our SLM in the top line of the chart. We can see that interest in GORE has risen significantly from 2008-12. The recent drop could be because of the nature of our Scopus citation cutoff, or could be a genuine drop in interest.

VI. Summary and Discussion

Survey Results. By analyzing the top-cited 246 papers as per Scopus, we've made several observations about the GORE field, enabling us to gain a high-level understanding of the progress made. We can observe some trends in research topics, notably a rise in adaptation/variability/evolution, but most of the popular topics seem to rise and fall with the number of

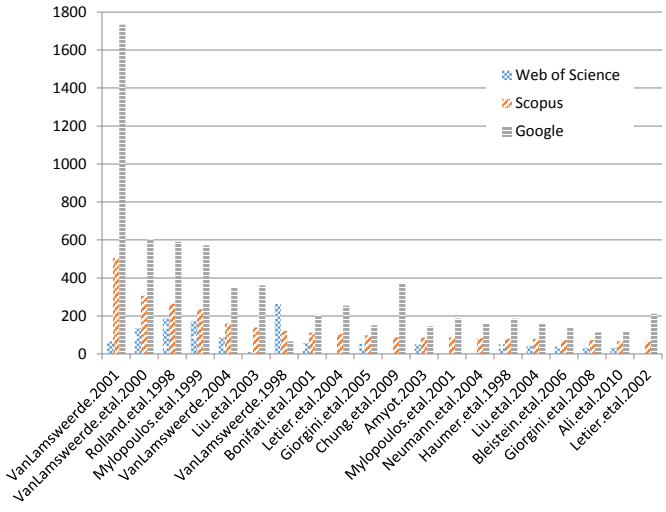

Fig. 8: Top 20 Cited GORE Publications Ranked According to Scopus Citations, also showing Google Scholar and Web of Science Citations (**RQ6**)

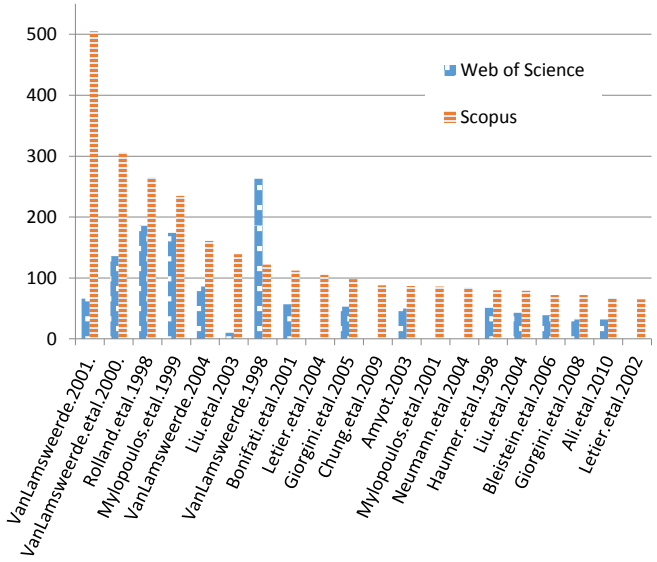

Fig. 9: Top 20 Cited GORE Publications Ranked According to Scopus Citations (Google Scholar Data Removed) (**RQ6**)

papers. The KAOS and i* Frameworks continue to dominate nearly equally, with the majority of papers remaining non-committal when it comes to selecting a particular framework. In terms of venue, RE and REJ dominate, but we can still see a wide variety of venues, with GORE-publications spread in many publication areas. Overall, GORE has seen increased interest in recent years, possibly with a dip in interest more recently.

We see that the focus of GORE historically was to propose new methods, but more extensive use of past approaches, in the form of implementations, integrations and extensions, appears to be rising slightly. In terms of evaluation, about half of the GORE publications contain a case study, with this number appearing to increase slightly; scalability tests are still in use, while other forms of evaluation are rare. Still, in conjunction with our findings of many papers with low citations, we would hope to see even more convergence and utilization of existing work, instead of a steady stream of new proposals.

We can hypothesize that the proliferation of new idea papers may be due to both the complex nature of many RE problems and the maturity of the field. RE, as a research field, is relatively new, and can be seen as a very rich research area with many difficult open problems that require complex new ideas as solutions. We can compare the scope and history of the field to other technical research areas. For example, Databases can be seen to be narrower and more focused, and after a few new ideas papers by Codd [5], successive work has been largely evaluation, application and innovation of industrial practice. On the other hand, AI is even broader than RE and so new ideas papers keep being produced. But it is also more mature in that there are more evaluations and applications to practice. Notice that both AI and Databases are more than 20 years older than RE.

In terms of complexity, we believe that new ideas which are more complex, addressing harder problems are more likely to see extensions. It can be argued that understanding and evaluating the socio-technical divide between complex human organizations and complex systems is a particularly hard problem, which may be why the area of GORE research appears to have difficulty converging.

It is interesting to compare the trends in GORE to other RE research topics. Although this is challenging without similarly structured SLMs for other topics, we are able to make some comparison regarding the use of empirical methods. An editorial by Daneva et al. evaluates the status of empirical methods in RE by looking at existing SLRs [6], finds that the number of empirical studies published in RE-venues has increased dramatically in the last ten years. In our results, although we note what appears to be a slight increase in Case Studies, we find the number of GORE-related papers with evaluation has increased more-or-less at the same rate of increase of GORE papers in general. In this light, GORE evaluation is increasing, but only relatively to the number of GORE publications. It is not clear whether this is also the case for RE in general, i.e., more evaluation studies because of more published papers in general.

For those planning on making future research contributions to GORE, we can use our data and analysis to make recommendations. 1) As the breadth of available GORE research is wide, due diligence is required to find related work. It's likely not enough to cite the "usual suspects", but a more detailed literature search should be performed; making an effort to understand, adapt, extend and re-use what has been done, instead of producing new proposals which may have a high-degree of overlap with existing work. 2) Plain clear wording in the title, abstract and keywords are important; both to be included in this and future Meta Studies, but also to help future readers more easily pick up on your work. 3) It would be ideal to see an increased focus on evaluation of existing methods, rather than the introduction of new ones.

Survey Method. Our initial conceptualization of the survey design was a "next generation survey", focusing on the evolution and maturity of particular ideas, going beyond the current state-of-the-art in Empirical Software and Requirements Engineering following the systematic process laid out by Kitchenham et al. Maturity could be defined by phases similar to those reflected in our paper types (proposal, formalization, implementation, evaluation, etc.), but the real measure of maturity could be gained by looking at the references for each paper to determine how the reference relates to the current work: does the paper extend, evaluate, map to, implement, formalize, etc. the related work, or does it just mention the other paper for the purposes of comparison, rather than an advancement of ideas?

We set the initial design of our survey along these lines, creating a classification scheme for references between papers. However, after two rounds of development and ICR testing, we found that agreement amongst coders for these reference classifications was too low, with an average of 0.14 after round two, far lower than our level of agreement for paper types and topics, even though the classification schemes for paper types and references are similar. These results in and of themselves are interesting. This tells us that 1) authors are not particularly good at making the relationship between their own work and other publications clear, 2) coders, even after much practice and discussion, are not sufficiently good at being able to classify the relationships between papers accurately. Furthermore, the process of coding references was incredibly time consuming, explaining why our ICR process of coding 55 papers took six months, while coding about 60 papers, twice, without considering references took us only about two months.

Related to this idea, the initial survey design had us using snowballing as the primary means of finding papers, in line with the focus on references and idea evolution in the next generation survey. Once we decided to abandon the next generation survey design and focus on a traditional systematic literature map, we adjusted our design to find candidate papers via systematic search.

VII. Threats to Validity

We can identify several threats to the validity of our study. Although we have covered 246 papers, we have omitted those papers with Scopus references ≤ 3 and workshop papers, threatening *Conclusion Validity*. By omitting workshop papers, we may miss influential work in the area, or work published in workshops which later became conferences. Given that this is a first general mapping of goal model work, we focus on those publications which are more rigorously peer-reviewed. Furthermore, by using the citation cutoff of three, we put greater emphasis on older work, discounting new work in the area which may not yet be extensively cited.

The inclusion or exclusion of papers in our survey may be subjective or error prone, i.e. does a paper involve GORE significantly? However, we mitigate this threat by having two people check the inclusion/exclusion of papers, and by discussing borderline papers.

It would have been desirable to have a more extensive set of topics. After creating the initial set of topics we found several further topics which were candidates for inclusion (e.g., Scenarios). A few of these topics were included after discussion and an agreed-upon definition; however, we avoided adding many new topics as we had already conducted our ICR tests, and could not guarantee the reliability of further topics.

Our systematic search criteria may also be subject to critique, threatening *Construct Validity*. As many scientific papers talk about the "goal" of the paper and have some sort of scientific model, searching for variants of "goal model" provides too many false positives. Likewise with synonyms of "goal", e.g., intention, motivation. Thus we chose to use only "goal" and to include "requirements" in our search. However, "goal" and "requirements" produces many papers that have nothing to do with modeling, thus we searched for variations of "goal model" and "requirements". We also experimented with the use of "engineering" in our query, but found the results too narrow. Although we arrived at what we believed

was an effective search string, we may miss papers which concern GORE but do not use variations of "goal model" in the title, abstract and keywords. A notable case is Yu's RE'97 paper [32], winning a most influential paper in RE'07, but not caught by our systematic search string as it does not explicitly use the terminology in our search string. This emphasizes the importance of keywords when writing papers, although in the case of [32] the terminology for GORE had not yet converged to recognizable keywords. Future work could expand our data using snowballing.

We have discussed extensively our coding process and measurement of ICR, relating to the *Internal Validity* of our results. We hoped that our ICR measure would be higher, particularly for paper topics, but given the challenges of qualitative coding, we accept these results. Given the large number of coders we had and the large number of categories, it was particularly challenging to achieve high agreement scores.

The authors of this study have significant experience in goal modeling (typically *i**-related languages). This may bias survey tagging, threatening *External Validity*. Several authors of this study are authors of publications included in the SLM. However, as candidate publications were found via systematic search and objective citation data, we mitigate most threats. Fig 6 shows the inclusion of many frameworks beyond i*.

Relating to External Validity, with any systematic literature review, it's important to demonstrate sufficient *Repeatability*. If another set of people were to reclassify the same group of publications using our tags, we have confidence that our tag definitions would help them make choices which are fairly consistent with our results. However, outsiders could not be present for our extensive discussions, and it is not feasible to make collected group knowledge explicit. Furthermore, if another group went through a different process of grounded paper topic building, as described in Sec. IV-A, they may come up with a different set of topics. This is an unavoidable side-effect of qualitative coding; nevertheless, we believe our results provide a useful contribution to the RE community. We make all of our survey data publicly available and welcome further analysis, including alternative codes.

VIII. Related Work

Literature Reviews in SE. We have created our roadmap by adopting the methods and approaches prescribed by Petersen et al. [23], specifically focusing on a map of available work, rather than a detailed survey evaluating publication quality, clearly defining our process of finding and including papers, making our research questions clear. As our survey is designed as SLM and not an SLR, we do not perform a deeper evaluation of paper quality, for example using criteria provided by Ivarsson and Gorschek [14]. In the trade-off between paper volume and survey depth, in choosing to conduct a SLM, we focus on volume, covering many papers in a shallow way. Future work should evaluate GORE literature, likely covering smaller sets of publications, in more depth.

Kitchenham et al. provide guidelines for empirical studies in software engineering. When applicable, we apply many of these guidelines to our systematic mapping study, including clearly specifying a hypothesis (in our case research questions), defining populations (publications from systematic search of Scopus), defining a process, providing raw data, and making extensive use of graphics [16], [18].

Work by Pham et al. focuses on a social network analysis of computer science publications, investigating collaboration and citations [24], applying such analysis to the CAiSE conference series in [15]. This would be an interesting way to evaluate the nature of the GORE community: is it interdisciplinary, hierarchical, etc, but we omit such an analysis in this paper due to space restrictions. Future efforts could use our data for this type of analysis.

Meta-Reviews in RE. Bano et al. perform a meta-review of systematic literature reviews in RE, finding that the number of systematic literature reviews in RE has increased dramatically from 2006 to 2014, but that the quality of such studies has decreased. They measure quality by looking at inclusion/exclusion criteria, search space adequacy, quality assessment of primary studies, and information regarding primary studies. A systematic literature map, by nature, does not evaluate the quality of or provide specific information regarding primary sources [18], [23], so we believe the latter two quality categories do not apply to this study (or other SLMs in RE). Regarding the first two points, we listed our inclusion/exclusion criteria in Sec. III and have selected Scopus as our publication source as it covers major databases in our field (IEEE, Springer, ACM) avoiding the need to combine results of multiple databases.

The Daneva et al. evaluation, looking at existing SLRs in RE [6], finds two reviews related to GORE [12], [8]. The former is the previous work of some of the authors, while the latter, focusing on compliance, was omitted from our survey as it appeared in a workshop.

GORE Literature Reviews. Our survey found 21 papers marked as Meta-Studies. From these studies, we cover the most general literature reviews in this section, with further GORE-related literature reviews such as [22] or [30], focusing on specific sub-topics such as law or reasoning. We find that most of the prominent GORE-related literature reviews were not performed systematically, with a few exceptions.

The most cited GORE literature review (also the most cited GORE paper) is van Lamsweerde's guided tour of the area as per 2001 [28]. This work motivates the use of goal-orientation and summarizes existing methods for modeling, specifying, and reasoning over goals. Chung & Leite review the state-of-the-art in Non-Functional Requirements, exploring definitions, classifications and representations of NFRs, reviewing prominent publications at the time [4]. The paper associated with van Lamsweerde's RE'04 keynote [29], provides an overview of work relating primarily to the KAOS framework.

In [1], Amyot and Mussbacher perform a SLR of publications, finding 281 publications using the User Requirements Notation (containing the Goal-oriented Requirement Language (GRL)). The focus of our current survey is broader and more shallow, looking at all GORE notations, including GRL, and

not getting into extensive details. In addition to presenting the goal/strategy map, Rolland and Salinesi perform an extensive overview of GORE as per 2005 [26]. Grau et al. compare and contrast six dialects of i* [9], while Regev and Wegmann review GORE methods in order to improve definitions of goal types using principles from regulation mechanisms [25].

As described in Sec. I, our previous SLM and SLRs have focused on GORE publications describing transformations/mappings, in order to understand how goal models can lead to downstream development [12], [13]. Although the RQs and inclusion/exclusion criteria of these publications bear similarities to the current work, the set of papers reviewed is different, as is the method used for finding the papers (systematic search with a citation cutoff vs. snowballing and systematic search with no citation cutoff).

IX. Conclusions and Future Work

We have provided the first general systematic survey of GORE, covering the 246 most cited publications, according to Scopus. We have chosen to give a high-level overview of the field using a SLM, with an emphasis on descriptive graphics. We have focused our inquiries with a number of specific research questions, and used our results to make general recommendations for future GORE-related research. In the name of repeatability and enhancing the knowledge of the field, we have made our publication data and category descriptions publicly available[5]. We encourage further analysis, investigation and expansion of our data.

We remain interested in the next generation survey concept, focusing on categorizing the relationships between publications to track the evolution of ideas. However, the difficultly in reliably tagging references between papers needs to be worked out, perhaps through some form of cooperative open tagging. Future work should investigate this and other possibilities.

Acknowledgments

This work is supported by: ERC advanced grant 267856, "Lucretius: Foundations for Software Evolution", http://www.lucretius.eu, an ERC Marie Skodowska-Curie Intra European Fellowship (PIEF-GA-2013-627489), a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (Sept. 2014 - Aug. 2016), by the European Union's Horizon 2020 research and innovation programme under grant agreement No 653642 - VisiON, and by the APOSTD grant (APOSTD/2014/064) from the Generalitat Valenciana. We thank Trento students who contributed to early stages of the SLM, and Dr. S. Stumpf for help with ICR.

References

[1] D. Amyot and G. Mussbacher, "User requirements notation: the first ten years, the next ten years," *Journal of Software*, vol. 6, no. 5, pp. 747–768, 2011.

[2] A. Anton and C. Potts, "The use of goals to surface requirements for evolving systems," in *ICSE*, 1998, pp. 157–166.

[3] M. Bano, D. Zowghi, and N. Ikram, "Systematic reviews in requirements engineering: A tertiary study," in *EmpiRE*, 2014, pp. 9–16.

[4] L. Chung and J. C. S. do Prado Leite, "On non-functional requirements in software engineering," in *Conceptual modeling: Foundations and applications*. Springer, 2009, pp. 363–379.

[5] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, Jun. 1970. [Online]. Available: http://doi.acm.org/10.1145/362384.362685

[6] M. Daneva, D. Damian, A. Marchetto, and O. Pastor, "Empirical research methodologies and studies in requirements engineering: How far did we come?" *J Syst Software*, vol. 95, pp. 1–9, 2014.

[7] A. Dardenne, A. Lamsweerde, and S. Fickas, "Goal-directed requirements acquisition," *Sci Comput Program*, vol. 20, pp. 3–50, 1993.

[8] S. Ghanavati, D. Amyot, and L. Peyton, "A systematic review of goal-oriented requirements management frameworks for business process compliance," in *RELAW*, 2011, pp. 25–34.

[9] G. Grau, C. Cares, X. Franch, and F. Navarrete, "A comparative analysis of i* agent-oriented modelling techniques." in *SEKE*, 2006, pp. 657–663.

[10] D. Harel and B. Rumpe, "Meaningful modeling: What's the semantics of "semantics"?" *Computer*, vol. 37, no. 10, pp. 64–72, Oct. 2004.

[11] HitachiGroup. (2016) Pentaho. Accessed: 2016-06-27. [Online]. Available: http://community.pentaho.com/

[12] J. Horkoff, T. Li, F.-L. Li, M. Salnitri, E. Cardoso, P. Giorgini, J. Mylopoulos, and J. Pimentel, "Taking goal models downstream: a systematic roadmap," in *RCIS*, 2014, pp. 1–12.

[13] J. Horkoff, T. Li, F.-L. Li, M. Salnitri, E. Cardoso, P. Giorgini, and J. Mylopoulos, "Using goal models downstream: A systematic roadmap and literature review," *IJISMD*, vol. 6, no. 2, pp. 1–42, 2015.

[14] M. Ivarsson and T. Gorschek, "A method for evaluating rigor and industrial relevance of technology evaluations," *Empirical Software Engineering*, vol. 16, no. 3, pp. 365–395, 2011.

[15] M. Jarke, M. Pham, and R. Klamma, "Evolution of the caise author community: A social network analysis," in *Seminal Contributions to Information Systems Engineering*, 2013, pp. 15–33.

[16] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *TSE*, vol. 28, no. 8, pp. 721–734, 2002.

[17] B. Kitchenham, T. Dyba, and M. Jorgensen, "Evidence-based software engineering," in *ICSE*, 2004, pp. 273–281.

[18] B. A. Kitchenham, D. Budgen, and O. P. Brereton, "Using mapping studies as the basis for further research – a participant-observer case study," *Inform Software Tech*, vol. 53, no. 6, pp. 638 – 651, 2011.

[19] K. Krippendorff, "Reliability in content analysis," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[20] J.-N. Mazón and J. Trujillo, "An mda approach for the development of data warehouses," *Decis Support Syst*, vol. 45, no. 1, pp. 41–58, 2008.

[21] Microsoft. (2016) Power bi. Accessed: 2016-06-27. [Online]. Available: https://powerbi.microsoft.com/en-us/

[22] P. N. Otto and A. I. Antón, "Addressing legal requirements in requirements engineering," in *RE*, 2007, pp. 5–14.

[23] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *EASE*, 2008, pp. 68–77.

[24] M. Pham, R. Klamma, and M. Jarke, "Development of computer science disciplines: a social network analysis approach," *Social Network Analysis and Mining*, vol. 1, no. 4, pp. 321–340, 2011.

[25] G. Regev and A. Wegmann, "Where do goals come from: the underlying principles of goal-oriented requirements engineering," in *RE*, 2005, pp. 353–362.

[26] C. Rolland and C. Salinesi, "Modeling goals and reasoning with them," in *Engineering and Managing Software Requirements*, 2005, pp. 189–217.

[27] C. Seaman, "Qualitative methods in empirical studies of software engineering," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 557–572, 1999.

[28] A. van Lamsweerde, "Goal-oriented requirements engineering: A guided tour," in *RE*, 2001, pp. 249–262.

[29] ——, "Goal-oriented requirements enginering: a roundtrip from research to practice," in *RE*, 2004, pp. 4–7.

[30] ——, "Reasoning about alternative requirements options," in *Conceptual Modeling: Foundations and Applications*. Springer, 2009, pp. 380–397.

[31] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: a proposal and a discussion," *Requir Eng*, vol. 11, no. 1, pp. 102–107, 2006.

[32] E. Yu, "Towards modelling and reasoning support for early-phase requirements engineering," in *RE*, vol. 97, 1997, pp. 226–235.

[5] http://www.cs.toronto.edu/goreslm/