Understanding Semantic Search on Scientific Repositories:
Steps towards Meaningful Findability

Thiago Gottardi
Claudia Bauzer Medeiros
Julio Cesar Dos Reis

{gottardi,cmbm,jreis}@ic.unicamp.br

Instituto de Computação
Universidade Estadual de Campinas

Laboratory of Information Systems
Instituto de Computação
UNIVERSIDADE ESTADUAL DE CAMPINAS
UNICAMP
CENTER FOR COMPUTING IN ENGINEERING & SCIENCES
UNICAMP

Agenda

- Introduction
- Review Method
- Results
- Discussion and Open Challenges
- Related Works
- Conclusions and Ongoing Efforts

Read Further:

Section 1
Page 1

Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

Agenda

- **Introduction**
- Review Method
- Results
- Discussion and Open Challenges
- Related Works
- Conclusions and Ongoing Efforts

Read Further:

Section 1
Page 1

Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

Introduction

- Sharing of results:
  - A key enabler for Open Science[1].
  - Reuse of results.

Open Science

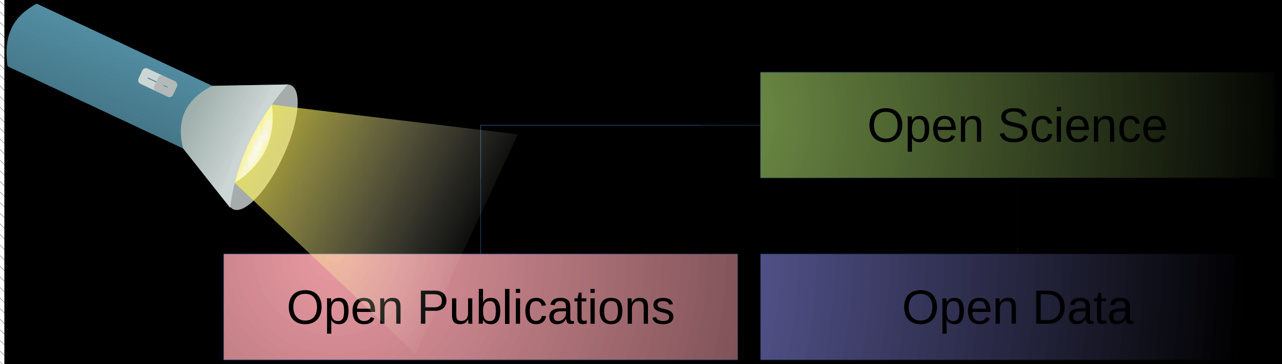
Axes: Open Publications | Open Data | Open Processes

[1] Woelfle, M., Olliaro, P., and Todd, M. H. (2011). Open science is a research accelerator. Nature Chemistry, 3:745–748.

Introduction

Problem:

Reuse depends on effective search mechanisms.



gottardi@ic.unicamp.br

Introduction

- Semantic search has been proposed for this issue:
  - Still, semantic mechanisms vary significantly.
  - Open questions remain:
    - What are the adequate mechanisms?
    - Which objectives and goals should be considered?
    - What data classes are searched?

Agenda

- Introduction
- **Review Method**
- Results
- Discussion and Open Challenges
- Related Works
- Conclusions and Ongoing Efforts

Read Further:

Section 2
Page 3

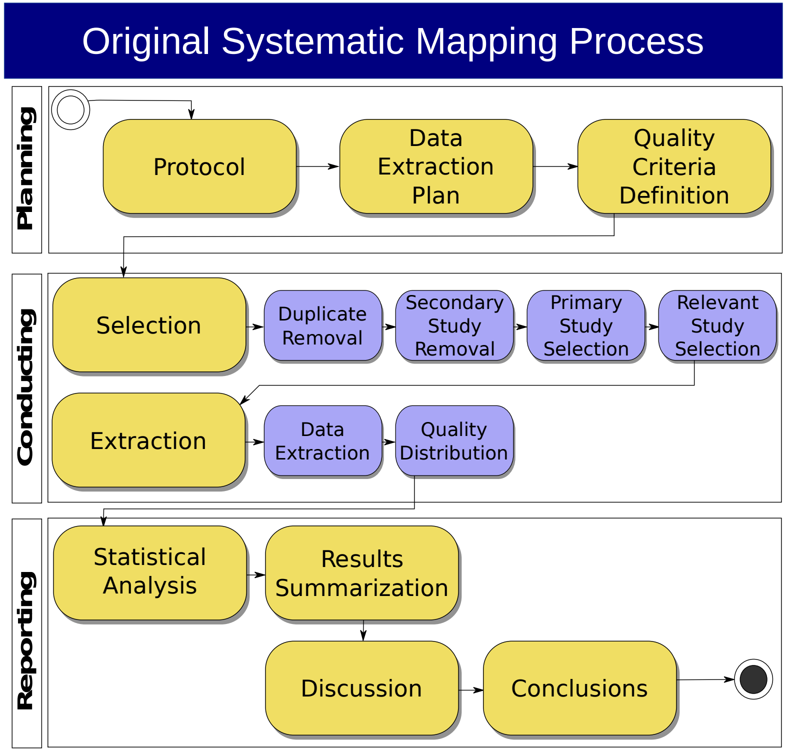
Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

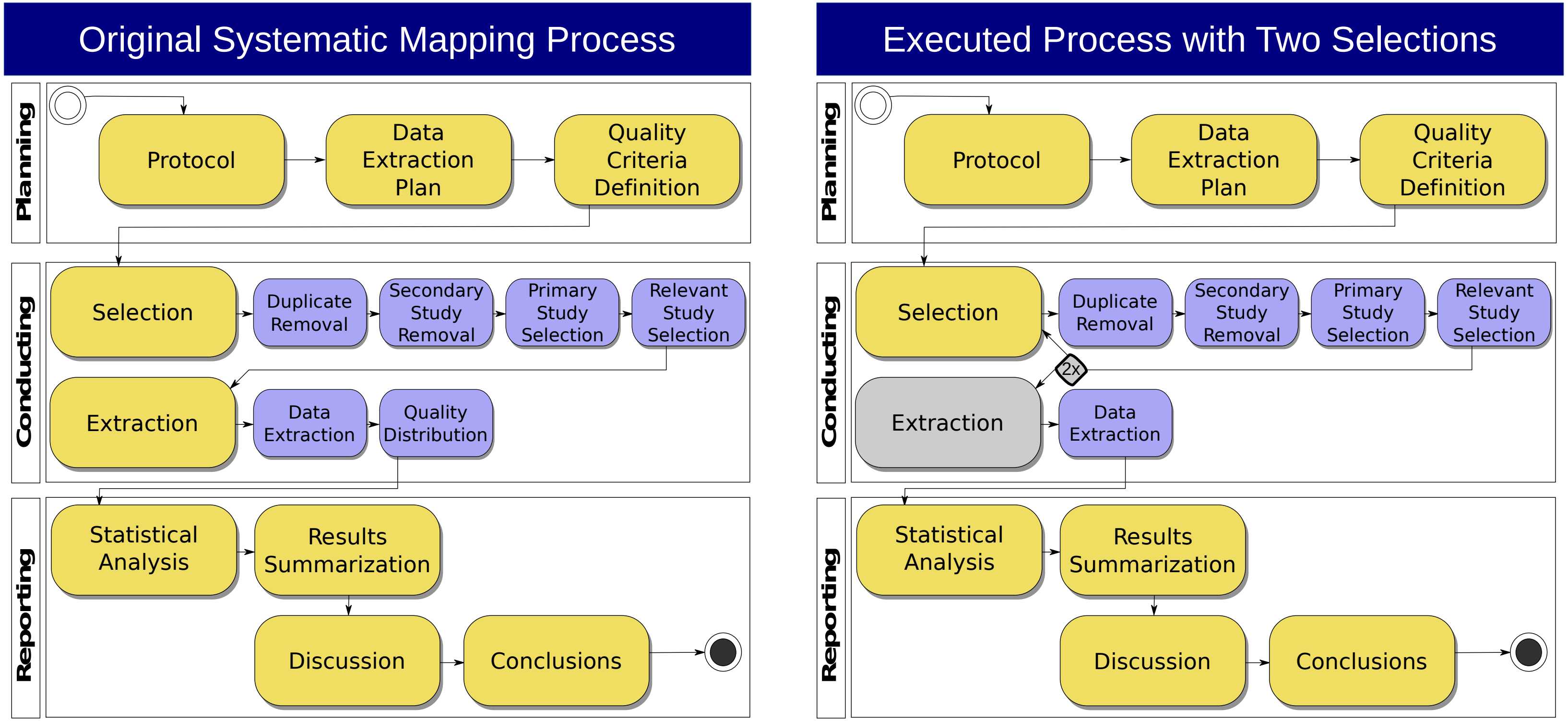
Method

- Systematic Mapping
  - A literature review based on a strict process;
  - Quantitative view on related publications.
    - Presents existing results and their numbers;
    - Lacks qualitative depth on their efficiency.

Review Process



Review Process



Review Protocol

Protocol Item	Item Description
Objective	Identify existing approaches to integrating semantic searches on scientific production.
Primary Research Question	RQ1: What are the approaches and techniques that perform integrated semantic searches on scientific production?
Secondary Research Question(s)	RQ2: What approaches or techniques employ semantic mapping? RQ3: What are the software architectures developed for integration? RQ4: What are the objectives for the proposal?
Intervention	Related primary studies must be identified and categorized.
Control	The search results must include previously known studies that are known by the researcher.

Review Protocol (Cont.)

Protocol Item	Item Description
Population	Search techniques and approaches.
Results (expected)	Quantitative data on approach frenquency distribution within scientific categories.
Application (expected)	Provided as a support to new research efforts.
Keywords	Semantic Search and Scientific.
Source selection criteria	Source must index studies on Computer Science, Mathematics or Engineering; must allow Boolean operators; must be accessible by the researchers.
Study Language(s)	At least title and abstract must be in English.
Search Engine(s)	Scopus and IEEExplore

Review Protocol (Cont.)

Protocol Item	Item Description
Selection Criteria	Inclusion: • (P1-101) I1 – Contains Search; • (P2-121) I2 – Integration or Semantic Mapping
	Exclusion: • (P1-1) E1 – Not a document or inaccessible; • (P1-2) E2 – Unrelated to computing/databases. • (P2-102) E3 – No search; • (P2-107) E4 – Not primary study*.

*Non primary studies must be verified for similarity prior to exclusion.

Search String Definition

Keyword	Synonyms
Semantic Search	"semantic search" ; "ontology search"; "metadata search"; "meta data search"
Search	“search”, “query”, ”information retrieval”, “retrieval”; ”access”
Scientific	"scientific"; "study pack"; "study packing"; ”research”

Search String Definition

Session	String
1	(("semantic search" OR "ontology search" OR "metadata search" OR "meta data search") AND ("scientific" OR "study pack" OR "study packing"))
2	(("semantic query" OR "ontology query" OR "metadata query" OR "meta data query") AND ("scientific" OR "study pack" OR "study packing"))
3	(("semantic information retrieval" OR "ontology information retrieval" OR "metadata information retrieval" OR "meta data information retrieval") AND ("scientific" OR "study pack" OR "study packing"))
4	(("semantic retrieval" OR "ontology retrieval" OR "metadata retrieval" OR "meta data retrieval") AND ("scientific" OR "study pack" OR "study packing"))

*Scopus included “Research” and “Science” for “Scientific”;
*Scopus included “Analogy” for “Semantic”;
*Scopus included “Retrieve” and “Access” for Retrieval;

Agenda

- Introduction
- Review Method
- **Results**
- Discussion and Open Challenges
- Related Works
- Conclusions and Ongoing Efforts

Read Further:

Section 3
Page 5

Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

Selection Phase 1

- Objective:
  - Select papers or articles;
    - Discard unrelated documents
      - e.g. talk reports, conference listings.
  - Select papers or articles related to search.

Phase Input	I1 (Search)	E1 (No Document)	E2 (Unrelated)	E3 (No Search)	Phase Output
299	280	9	1	4	276

Selection Phase 2

- Objective:
  - Select papers or articles related to integration.
    - Verify relevance of non-primary;
      - Discard non-primary studies.

Phase Input	I2.1 - Integration	I2.2 - Semantic Mapping	I2 – I2.1 ∩ I2.2 (Integration and Semantic)	I2 – I2.1 U I2.2 (Integration or Semantic)	E4 – Non Primary	Phase Output
276	82	20	12	90	8	85

Extraction Phase

- Objective:
  - Extract data according to the Research Questions.
    1) Semantic Search and its Integration;
    2) Semantic Mapping;
    3) Software Architectures;
    4) Information Usage Objective.
- Summarize selected documents.

Extraction Result

- Integration and Semantic Mapping:
  - 11 studies: {2007..2019}.

Best described on Paper:

Subsection 3.1
Page 6

Year	Type	Author	Title
2007	conference	Xiaoming, Z.	Material Scientific Data Integration for Semantic Grid
2008	conference	Pirrò, G.	Advanced semantic search and retrieval in a collaborative peer-to-peer system
2012	conference	Deus, H.F.	Translating standards into practice - One Semantic Web API for Gene Expression
2013	conference	Khattak, A. M.	Context-Aware Search in Dynamic Repositories of Digital Documents
2013	article	Luo, Y.	Dynamic mapping processing between global ontology and local ontologies in grid environment
2014	article	Abburu, S.	A generic mapping method and tool to execute semantic queries on relational database
2014	article	Zheng, S.	Enabling Ontology Based Semantic Queries in Biomedical Database Systems

Agenda

- Introduction
- Review Method
- Results
- **Discussion and Open Challenges**
- Related Works
- Conclusions and Ongoing Efforts

Read Further:

Section 4
Page 7

Browse extra documents and data:

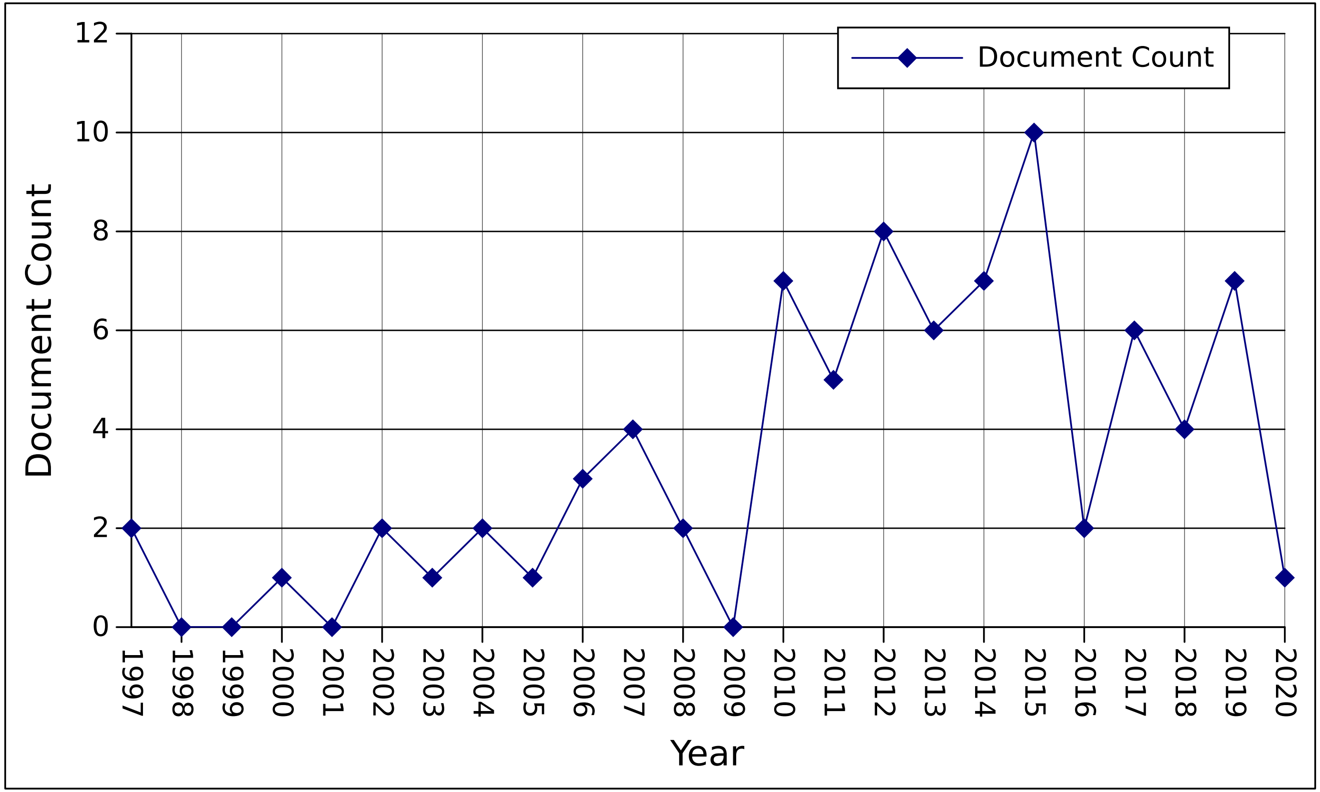
http://tiny.cc/gottardi-semantic-review

Semantic Search and Integration

- Semantic Search:
  - Depends on metadata and techniques to be efficient;
  - Different metadata formats and techniques were described.

- Integrated Search (77 studies):
  - Integration of different databases (39);
  - Integration of semantics to be added to existing data (34);
  - Integration of a semantic layer mapped to existing semantics (34);

Semantic Search



- Trend throughout the past decades.
- Note:
  - Recent years may receive more publications.

Semantic Search

- Studies that mention integrated semantic search:
  - 77 studies were found: {1997..2020}.

Year	Type	Author	Title
1997	article	Cardiff, J.	Semantic query processing in the venus environment
1997	article	Schatz, B.R.	Information retrieval in digital libraries: Bringing search to the net
2000	conference	Bukhres, O.	Effective standards for metadata in the GCMD data access system
2002	conference	Higgins, D.	Managing heterogeneous ecological data using Morpho
2002	conference	Nelson, C.	Use of metadata registries for searching for statistical data
2003	conference	Zhang	A practical approach for microscopy imaging data management (MIDM) in neuroscience
2004	conference	McClean, S.	MISSION: an agent-based system for semantic integration of heterogeneous distributed statistical information sources
2004	article	Yang, R.	Automatic metadata ingestion for supporting a web-based scientific

Semantic Search

- Broad areas:
  - These are provided as a rough and non-exhaustive guide.

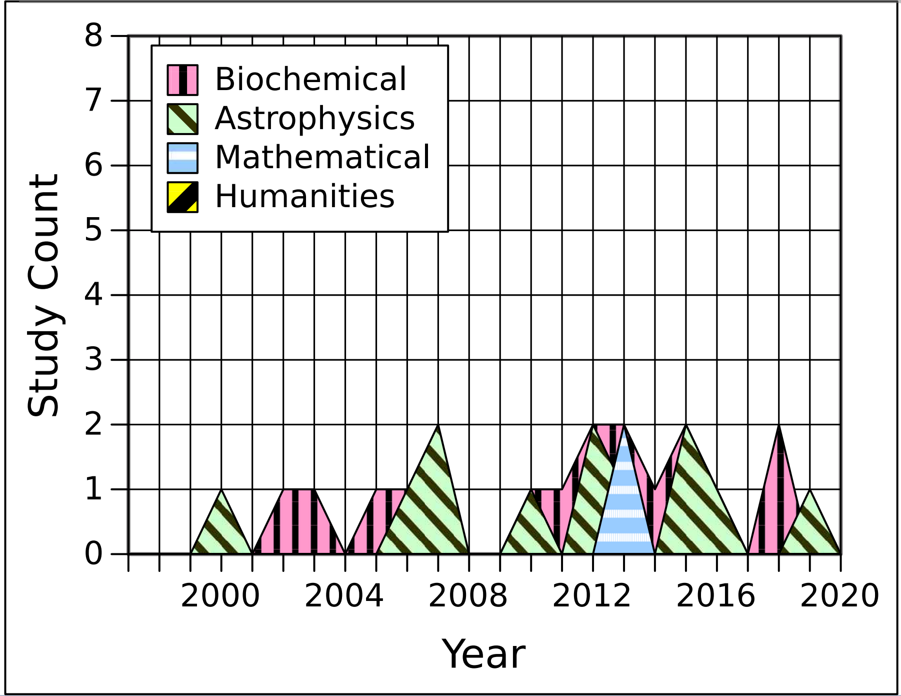
Biochemical, including:
Chemistry;
Biology;
Medicine.

Astrophysics, including:
Astronomy;
Physics;
Geology.

Mathematical, including:
Calculus;
Statistics;
Algorithms.

Humanities, including:
Cultural Heritage;
General Literature;
Human History.

Research Areas



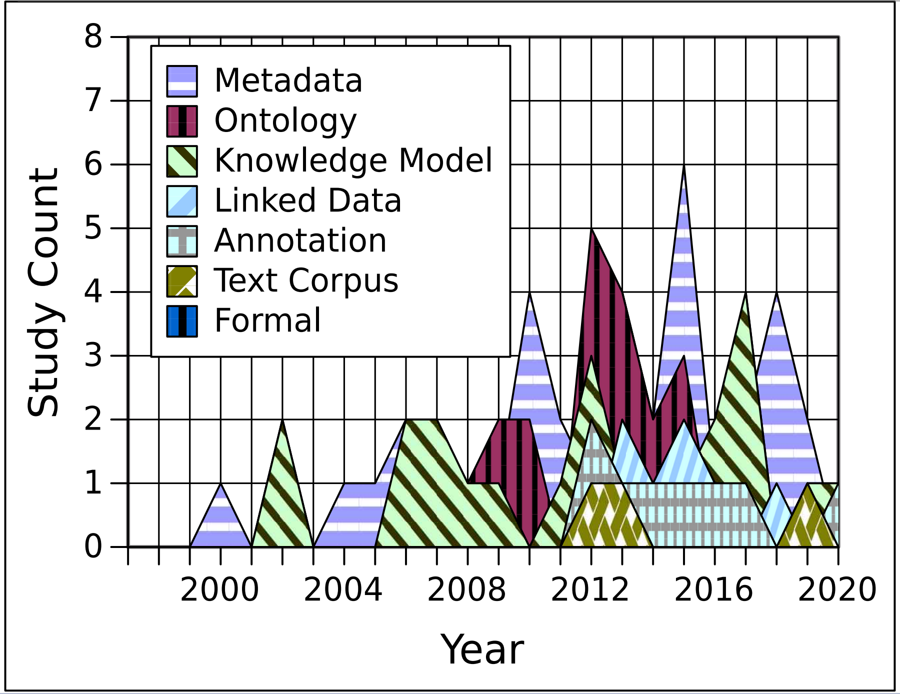
Semantic Search

- Most common metadata:
  - Presented as categories.
  - “Metadata” represents unclear.

Metadata includes unclear metadata.

Ontology also includes OWL and RDF.

Metadata Categories



Semantic Mapping

- Semantic layers have been proposed (17 studies):
  - Automatic (9); Manual (8); Fuzzy (3); Strict (0).

Automatic:
Computers process existing data
Algorithms identify and add metadata.

Manual:
Authors or curators work manually;
Humans manually add metadata.

Fuzzy:
Recommender systems use probability to suggest roughly adequate metadata.

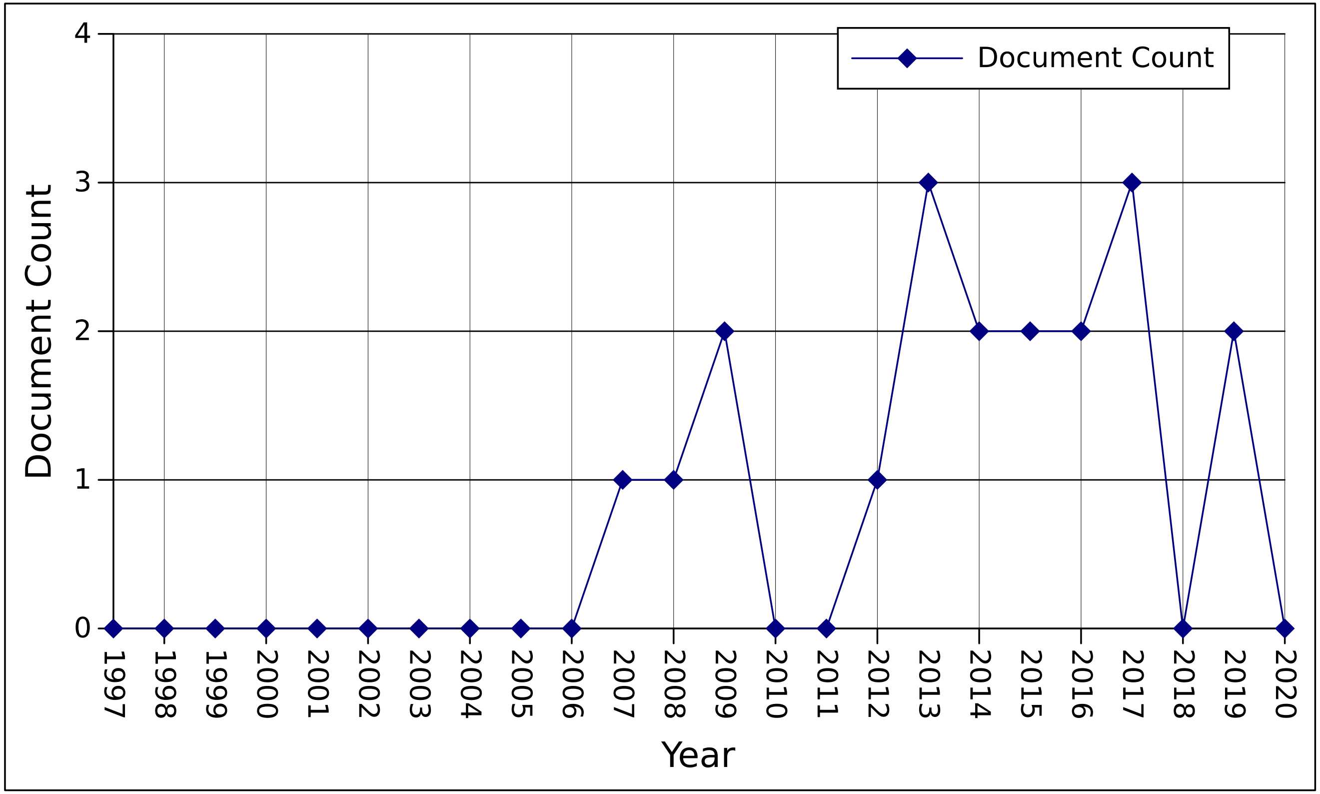
Strict:
Constraint rules enforce checks to ensure only correct metadata is added.

Semantic Mapping

- Studies that mention semantic mapping:
  - 18 studies were found: {2007..2017}.

Year	Type	Author	Title
2007	conference	Xiaoming, Z.	Material Scientific Data Integration for Semantic Grid
2008	conference	Pirrò, G.	Advanced semantic search and retrieval in a collaborative peer-to-peer system
2009	article	Liu, X.	Management of scientific principle knowledge for product innovation
2009	conference	Song, J.	Case study on multi-classifications based scientific data management and analysis system
2012	conference	Deus, H.F.	Translating standards into practice - One Semantic Web API for Gene Expression
2013	conference	Khattak, A. M.	Context-Aware Search in Dynamic Repositories of Digital Documents

Semantic Mapping



- Trend throughout the past decades.
- Note:
  - Recent years may receive more publications.

Semantic Search and Integration

- Different Meanings for Integration;
- Semantic Mapping Definitions:
  - Automatic; Manual; Fuzzy; Strict.
- Remaining challenge to balance
  - Domain-Specific and Generic.
- Different software architectures were described:
  - Major trend: migration from Clusters to Cloud.

Objectives and Data Classes

- Objectives:
  - Goals in which the data was originally stored or retrieved for.
    - Example: Manage data.

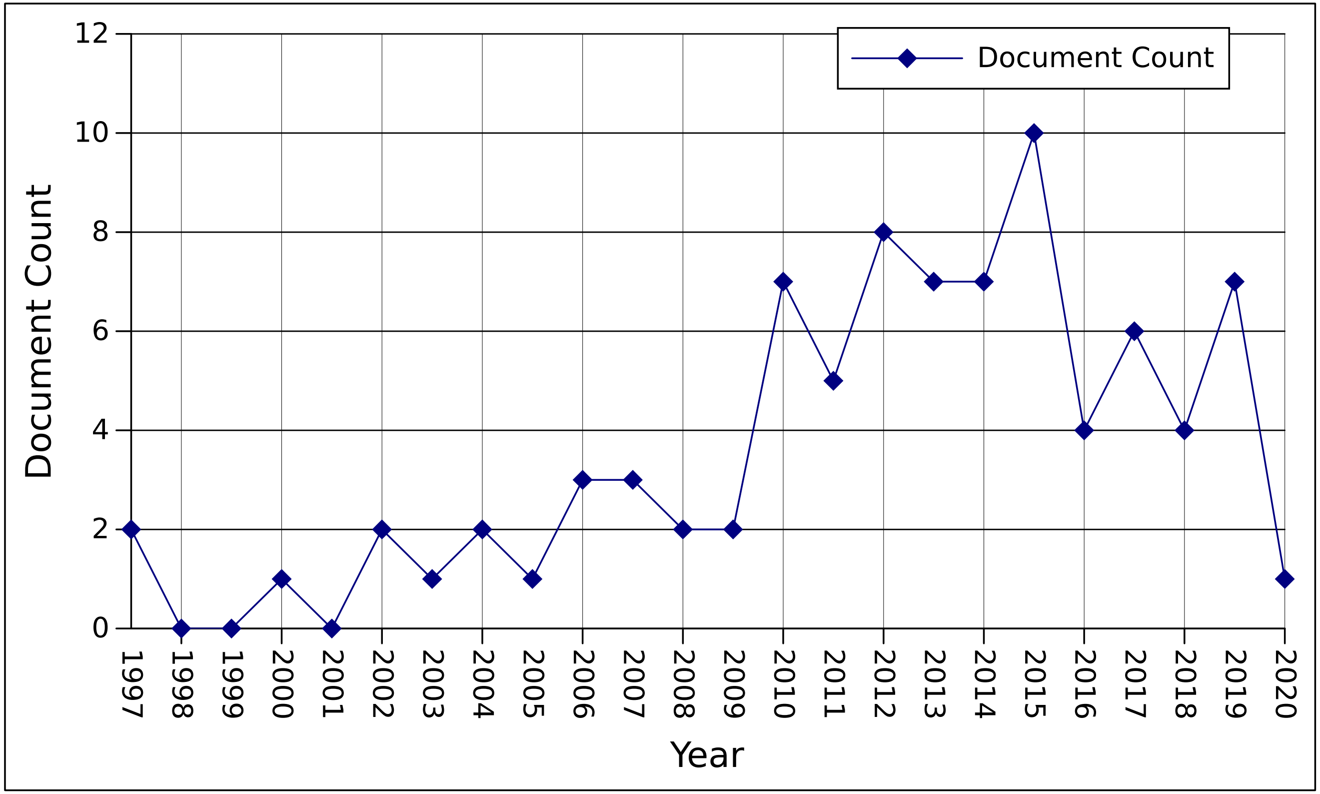
- Data Classes:
  - Category or datatype of stored data.
    - Example: Documents.

Objectives and Data Classes

- Studies that mention Objectives or Data Classes:
  - 85 studies were found (all): {1997..2020}.

Year	Type	Author	Title
1997	article	Cardiff, J.	Semantic query processing in the venus environment
1997	article	Schatz, B.R.	Information retrieval in digital libraries: Bringing search to the net
2000	conference	Bukhres, O.	Effective standards for metadata in the GCMD data access system
2002	conference	Higgins, D.	Managing heterogeneous ecological data using Morpho
2002	conference	Nelson, C.	Use of metadata registries for searching for statistical data
2003	conference	Zhang	A practical approach for microscopy imaging data management (MIDM) in neuroscience
2004	conference	McClean, S.	MISSION: an agent-based system for semantic integration of heterogeneous distributed statistical information sources
2004	article	Yang, R.	Automatic metadata ingestion for supporting a web-based scientific data and information super server

Objectives and Data Classes



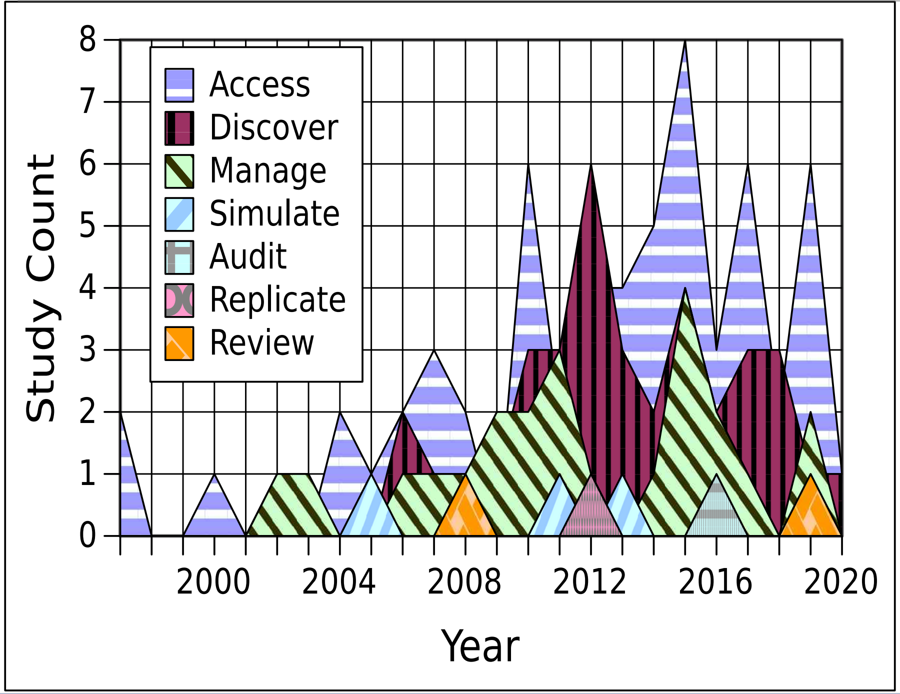
- Trend throughout the past decades.
- Note:
  - Recent years may receive more publications.

Objectives and Data Classes

- Most common objectives:
  - Presented as categories.
  - “Access” also includes search.

Access is usually combined with other usages.

Usage Objectives



Objectives and Data Classes

- Most common data classes:
  - Presented as categories.
  - Unclear classes are excluded.

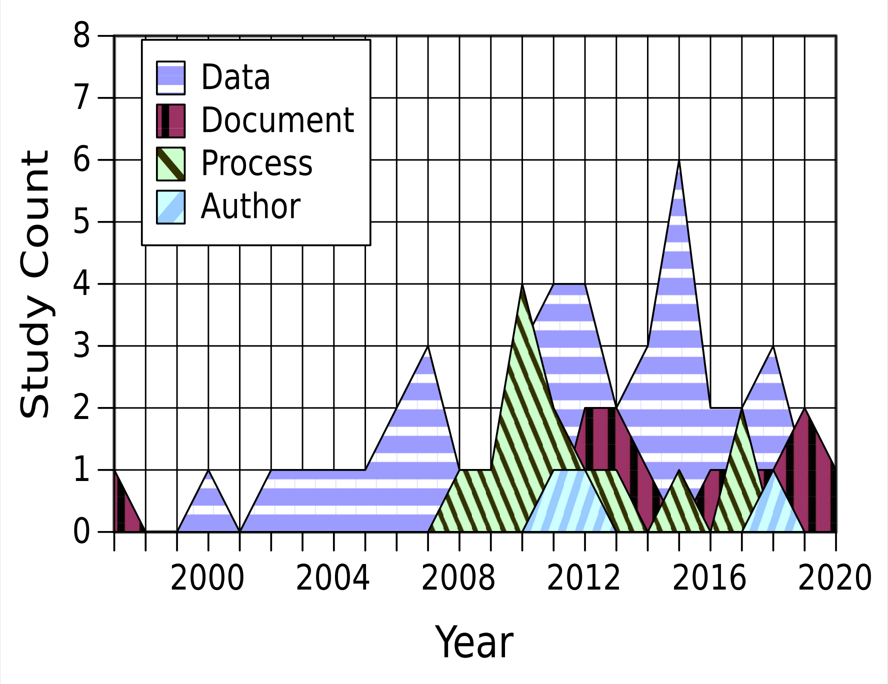
Data, including:
Numeric data;
Images;
Multimedia.

Document, including:
Papers;
Articles;
Reports.

Author, including:
Author Names;
Author Affiliation;
Research Groups.

Process, including:
Workflows;
Software;
Algorithms.

Data Classes



Objectives a

- Objectives x Data Classes:
  - Combinations indicate challenges and new opportunities.

Objective		Class: Scientific Data	Document	Process	Authors
	Access	29: Search, query, access, recommend and/or retrieve science data.	10: Search, query, access, recommend and/or retrieve papers, articles, journals, reports, magazines, etc.	8: Search, access, recommend and/or retrieve science data.	2: Search and find or recommend authors and related authors.
	Discover	22: Discover conclusions using aggregated science data.	4: Discover conclusions and related documents using existing documents.	7: Discover combined workflows.	1: Discover what authors collaborate on research efforts.
	Manage	13: Manage known science data, also their sources and bases.	2: Manage known document references/citations. Manage documents being written.	5: Manage known workflows and assess their usage.	1: Manage known authors, relationships, contributions and their roles.
	Simulate	3: Simulate experiments and compare against existing data for validation.	0: Simulate document publications and acceptance.	1: Simulate workflow usage and outcomes.	0: Simulate author contributions and outcomes.
	Audit	1: Audit data for validation and verification; protect from corruption and false data; blame manipulators.	0: Audit documents to verify authorship and protect documents from corruption.	0: Audit execution of workflows. Audit who can edit the workflow.	0: Audit roles and authorship to protect authors' curricula from corruption and false data.
	Replicate	1: Replicate studies based on existing science data and compare the outcomes.	0: Replicate (or plagiate) existing documents and their structures.	0: Replicate existing work-flows and compare their outcomes.	0: Plagiate author roles.
	Review	0: Review and compare data sets of science data to aggregate results.	1: Support for literature reviews.	0: Review work-flows and methods and compare their eff ciency.	0: Review existing author roles and contributions.

Objective	Science Data	Document	Authors	Process	Future Work	Call for Contributions and Research Topics
	Data Class					
Discovery	Discover conclusions using aggregated science data.	Discover conclusions and related documents using existing documents.	Discover what authors collaborate on research efforts.	Discover combined workflows.	Discover possible future works.	Discover trends for new topics and their calls for contributions.
Management	Manage known existing science data, also their sources and bases.	Manage known document references/citations. Manage documents being written.	Manage known authors, relationships, contributions and their roles.	Manage known workflows and assess their usage.	Manage possible future works.	Manage calls for conferences and their relationships.
Replication	Replicate studies based on existing science data and compare the outcomes.	Replicate (or plagiate) existing documents and their structures.	Plagiate author roles.	Replicate existing workflows and compare their outcomes.	Replicate goals for future works. Execute known future works.	Replicate interests from similar venues.
Review	Review and compare data sets of science data to aggregate results.	Literature reviews.	Review existing author roles and contributions.	Review workflows and methods and compare their efficiency.	Review past future works and compare against more recent past works.	Review calls from venues and compare their interests.
Simulation	Simulate experiments and compare against existing data for validation.	Simulate document publications and acceptance.	Simulate author contributions and outcomes.	Simulate workflow usage and outcomes.	Simulate future work outcomes prior to execution.	Simulate new trends for topics and calls for contributions.
Access (incl. semantic and recommender)	Search, query, access, recommend and/or retrieve science data.	Search, query, access, recommend and/or retrieve papers, articles, journals, reports.	Search and find or recommend authors and related authors.	Search, access, recommend and/or retrieve science data.	Search or recommend compatible past future works.	Search or recommend calls for contributions.
Audit	Audit data for validation and verification; protect from corruption and false data; blame manipulators.	Audit documents to verify authorship and protect documents from corruption.	Audit roles and authorship to protect authors' curricula from corruption and false data.	Audit correct execution of workflow. Audit who can edit the workflow.	Audit future execution of future work. Feasibility of future work.	Audit acceptance of venue according to the call for contributions.
Prediction	Estimate future science data production.	Estimate future document publications and demand. Estimate future document citations.	Estimate future authorship/contribution increase or decrease.	Predict outcomes for workflow usage/risk analysis.	Predict probability of execution for future work. Predict possible future works.	Predict possible future calls for contributions.
Strategic	Identify strategic data sets for future use.	Plan future documents to be written.	Plan roles for authors and contributions.	Establish new workflows. Plan workflow acceptance.	Plan current or future future works. Plan execution of future works.	Plan for future venue calls.
Public Visualiz. (not recommender)	Graphic views for aggregated scientific data.	Suggest relevant documents for public.	Suggest related authors for public.	Suggest workflows for public.	Show planned/expected future works.	Show expected/future calls for contribution.
Internal Access	(Easily) select filtered specific data/query within scientific databases.	Query internal text, sections, figures and tables from the documents.	Query author details.	Query workflow steps and roles.	Query details from future works.	Query interests and details from calls for contributions.
Result Export (exists within Access, but not declared)	Export scientific data.	Export Document searches and their results.	Export author names, affiliations and statistics.	Export workflows for reference and usage.	Export future work references.	Export venue calls.
[Graduate] Teaching (exists within Access, but not declared)	Select scientific data and adapt for teaching new classes or (post)graduate students.	Select adequate documents to elaborate teaching material.	Select relevant authors for students to study about.	Select relevant teaching methods or workflows to be studied.	Select relevant future works to be researched.	identify relevant topics to be taught.

Objectives and Data Classes

- Four Main Classes:
  - Data; Documents; Processes; Authors.
  - The first three are related to **Open Science** Axes.
- Processes and Software Repositories are related.
- Studies rarely employ more than one class;
  - Existing research challenge on combining classes.

Agenda

- Introduction
- Review Method
- Results
- Discussion and Open Challenges
- **Related Works**
- Conclusions and Ongoing Efforts

Read Further:

Section 1
Page 1

Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

Related Works

A survey of scientific metadata schema

Xu, H., Sun, L., Zou, M., and Meng, A. (2013)

Exploring metadata search essentials for scientific data management

Zhang, W., Byna, S., Niu, C., and Chen, Y. (2019)

Mapping a decade of linked data progress through co-word analysis

Niknia, M. and Mirtaheri, S. (2015)

The study of semantic and ontological features of thesaurus and ontology-based information retrieval systems

Karimi, E., Babaei, M., and Beheshti, M. (2019)

Agenda

- Introduction
- Review Method
- Results
- Discussion and Open Challenges
- Related Works
- **Conclusions and Ongoing Efforts**

Read Further:

Section 5
Page 11

Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

Ongoing Efforts

- Review Updates:
  - Further studies are being analyzed;
  - Update with Wiley Search was recently concluded.
    - ACM-DL and EV are under review.

- Data Sharing:
  - More data is planned to be shared and curated.

Browse extra documents and data:

http://tiny.cc/gottardi-semantic-review

Conclusions

- Integration and Mapping are referenced in different meanings;
- Semantic search infrastructures usually focus on a low number of data classes;
- Review support systems are not integrated with other objectives;
- New infrastructures should be planned to support future objectives and research fields requirements
  - Motivated by the balance between Domain-Specific and Generic.

References

Karimi, E., Babaei, M., and Beheshti, M. (2019). The study of semantic and ontological features of thesaurus and ontology-based information retrieval systems.
Iranian Journal of Information Processing and Management, 34(4):1579–1606.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.
Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, UK.

Niknia, M. and Mirtaheri, S. (2015). Mapping a decade of linked data progress through co-word analysis.Webology, 12(2).

Oakley, A., Gough, D., Oliver, S., and Thomas, J. (2005). The politics of evidence and methodology: lessons from the eppicentre. Evidence & Policy, 1(1):5–32.

Woelfle, M., Olliaro, P., and Todd, M. H. (2011). Open science is a research accelerator.
Nature Chemistry, 3:745–748.

Xu, H., Sun, L., Zou, M., and Meng, A. (2013). A survey of scientific metadata schema.
Applied Mechanics and Materials, 411-414:349–352.

Zhang, W., Byna, S., Niu, C., and Chen, Y. (2019). Exploring metadata search essentials for scientific data management. In 2019 IEEE HiPC, pages 83–92.

References

Karimi, E., Babaei, M., and Beheshti, M. (2019). The study of semantic and ontological features of thesaurus and ontology-based information retrieval systems.
Iranian Journal of Information Processing and Management, 34(4):1579–1606.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.
Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, UK.

Niknia, M. and Mirtaheri, S. (2015). Mapping a decade of linked data progress through co-word analysis.Webology, 12(2).

Oakley, A., Gough, D., Oliver, S., and Thomas, J. (2005). The politics of evidence and methodology: lessons from the eppicentre. Evidence & Policy, 1(1):5–32.

Woelfle, M., Olliaro, P., and Todd, M. H. (2011). Open science is a research accelerator.
Nature Chemistry, 3:745–748.

Xu, H., Sun, L., Zou, M., and Meng, A. (2013). A survey of scientific metadata schema.
Applied Mechanics and Materials, 411-414:349–352.

Zhang, W., Byna, S., Niu, C., and Chen, Y. (2019). Exploring metadata search essentials for scientific data management. In 2019 IEEE H

+ 299 papers and articles available as packing.

Acknowledgements

Thanks in advance for your great questions.

Thiago Gottardi, Claudia Medeiros *and* Julio dos Reis

FAPESP
Fundação de Amparo à Pesquisa do Estado de São Paulo

2019/19389-1
2017/02325-5
2013/08293-7

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

428459/2018-8
305110/2016-0

Thank you

For useful reviews.

For your attention.