TIB | German National Library of Science and Technology

The Quest for Research Information

Ina Blümel, Lambert Heller, Martin Mehlberg @ *TIB Hannover*
Stefan Dietze, Robert Jäschke @ *L3S Research Center, LU Hannover*

Dr. Ina Blümel
CRIS 2014 conference
May 13, 2014

TIB|Open Science Lab

Context

- New research network “Science 2.0” of german Leibniz Association
- New department “Research & Development” at TIB Hannover

Actually focusing on **2 research areas**

1. Writing and publishing quality-assured academic literature on a collaborative platform. Booksprint #CoScience
2. Utilizing the potential offered by CRIS by publishing standardised, open data
   - available information across institutions for entire scientific communities that build on this data
   - emerging through standards and tools such as VIVO application, ontology, …

Current activities in context of CRIS

- „VIVO for scientific communities“ implementations
  - E.g. a prototype for harvesting and structuring research information, student project at HSH
- Community building
  - VIVOcamp13 (first workshop for EU VIVO community, SWIB13 satellite)
  - VIVO camp14 at ELAG Conference (European Library Automation Group), 10-13 June 2014, Bath; more „hands-on“
- Policy & Standards making: Position paper DINI AG FIS
- DFG application: “German Academic Web”

Paradigm shift

- Not about institutional view on RI
  - research management as a driving force: reporting, financial/project management, etc.
  - and their (mostly proprietary) CRIS implementations at the institutional and partly even national level, …

- But rather **cross-institutional** view
  - Added value by **merging & linking research information from various sources**
  - Thinking outside the box, interdisciplinary manner, establish networks, …(see success of ResearchGate, academia.edu, etc.) Researchers and their **scientific communities**

Use cases and examples - Quests like ..

Which universities focus is on the geosciences?

Scholars
writing research proposals

Which articles have been accepted or distinguished by the relevant conferences?

…updating about actors
involved in subject area

What is the structure of cooperative activities (publications / projects) between research institutions?

analyses in science research

Which work groups are working at the interface between computer science and biology?

…establishing networks
with colleagues

How does the institute/staff structure of engineering institutes differ from that in natural science institutes?

Which industrial partners or public bodies are involved in e-learning projects?

Public

Need for Research Information

Who?

Stakeholders like researchers, research managers and administrators, policy makers, research councils and technology transfer organisations, the media and the general public (context “citizen science”)

How?

- Consistent
  - + temporal aspects, like past engagements of researchers, are captured
  - + persistent identifiers like ORCID are integrated
- Up to date
- openly available and reusable

Review: Sources for Research Information

source	example	data source	temporal aspects, like past engagements of researchers, are captured	data is international and/or cross-institutional	retrieval for specific disciplines or institutions is almost complete	integration of persistent indentifiers like ORCID implemented or planned	data is reusable
A. literature database	PubMed, Scopus, DBLP, arXiv, CiteSeer	publisher, professional associations, etc.	*	*	yes	yes	**
B. search engine	Google Scholar, Microsoft Academic Search	web crawling (A, C, D, F)	no	no	*	yes	**
C. individual/ institutional website		manual input by researchers, often augmented by A	yes	no	no	no	yes
D. social network	Academia.edu, Mendeley, ResearchGate	manual input by researchers, often augmented by A	*	no	no	yes	*
E. Wikipedia/ Wikidata		manual input by Wikipedia authors	yes	yes	no	yes	yes
F. research information system	Atira Pure, Avedas Converis	master data of research facilities, often manually complemented by researchers and by A	*	yes	yes	no	yes
G. VIVO aggregator	VIVO Search, CTSA Search, AgriVIVO	pure aggregation from F	yes	yes	yes	yes	no

** In some (of the here mentioned) cases possible, but not always.*

*** Does not apply, because publication metadata is prevalent.*

TIB

Current Research Information

- No holistic view of scientific „landscapes“
- No substantial quantities of freely available data
- Available data: published in different formats on different sites
- Maintaining research information: a burdensome task which ties up resources

From the vast array of research objects on the web to browsing in complete & linked researcher profiles!

How get the data?

Data Capturing

(1) Manually

- □ High effort
- □ Many redundancies
- Successful when in the self-interest of researchers
  → Benefit: crowdsourcing, certain amount of control

(2) Automatically

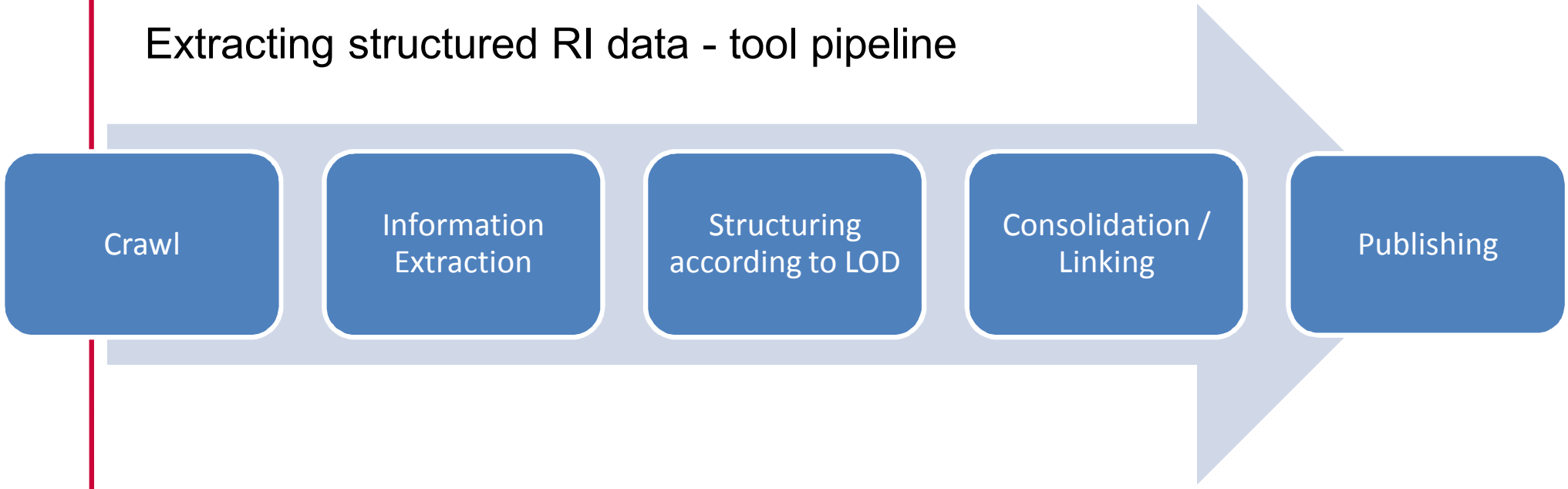
- *Structured* data that can be captured automatically
- □ Rarely standardised interfaces (e.g. DBLP), often access restrictions

(3) Alternatively

- *Unstructured* data is captured with *web crawlers* (e.g. Heritrix)
- Challenge: *focused crawling* of relevant web pages (referring to domain, file type, certain topic - matter of ongoing research; identification of research pages using machine learning approaches)

After crawling unstructured information

Extracting structured RI data - tool pipeline



- Entity recognition major task in information extraction (applied in e.g. ontology generation, text classification)
- Consolidation with authority files + linking of data: improve coherence and richness of automatically extracted data (also using background knowledge from related structured data sources)

Goals

Holistic approach for automated creation of scientific information

Representation of the research landscape in the Linked Open Data Cloud

- Application e.g.
  - Bootstrapping building (CERIF-based) CRIS
  - Community VIVOs

- Supporting scientists: devising new research issues, searching cooperation partners + research projects
- Tapping additional sources of alternative metrics for assessing the significance of scientific results
- Supporting exploration of different facets and complex enquiries and analyses
- Facilitating the research information maintainance
- Facilitating a transparent, evidence-based scientific policy

TIB | German National Library of Science and Technology

Thank you for your attention!