Open science and research assessment
Which agenda should the informatics community push forward?

Pierre Paradinas, SIF & Cnam
Laurent Romary, Inria
Why should we care about open science and evaluation?

From DORA to COARA

- the quick (and positive?) evolution of the research assessment context

Open science

- more open content available for (possibly?) monitoring research

Which role for the informatics community?

- expressing our needs
- setting up a political and technical roadmap
- specific context of the EU GraspOS project: CS pilot (Inria, UniBO, Athena)
The evolution of assessment practices

Towards more qualitative oriented assessment of research

● DORA declaration - [https://sfdora.org](https://sfdora.org) - 2013
  ○ “To call attention to **new tools and processes in research assessment** and the **responsible use of metrics** that align with core academic values and promote consistency and transparency in decision-making”

● CoARA (Coalition for Advancing Research Assessment) - december 2022
  ○ Universities, RPO, RFO, assessment organisation, infrastructures
  ○ Taking a wider range of objects into account
    ■ “the full range of research outputs, such as scientific publications, data software, models, methods, theories, algorithms, protocols, workflows, exhibitions, strategies, policy contributions, etc.”
  ○ Participation of various ICT-related institutions: Inria, Inesc TEC, IBICT, GRIN, SCIE, DARIAH
The accelerating open science agenda

From open access to open science

- publications
  - issues related to APC, diamond model, text and data mining possibilities, sovereignty at large (who has the corpus?)
- data sets
  - more constrained than publications (sensibility, rights, formats, communities)
- software
  - a beast of its own: specific issues related to authorship, versioning, reproducibility, relation to data

Open science policies: institutions, countries, Europe

- e.g. in France: national open science plan (V2, July 2022)

Infrastructures for open science

- e.g. in France: HAL, Recherche Data Gouv, Software Heritage
The specificities of scholarly practices in informatics

Central role of software

- Corresponding assessment practices:
  - Évaluation des Logiciels (Assessing software), Canteaut et al. 2021:

Conferences and workshops:

- More publications than in journals
- Quality peer-review => significant weight in assessment practices
- Strong variations across communities
  - https://github.com/societe-informatique-de-france/referentiel-pratiques-publication-2019
- Difficulty in multidisciplinary assessment contexts, hence:
  - low representation in commercial bibliographic databases (Kuserow & Groppe, 2014)
  - lack of authority data for conferences, which is also a difficult topic (naming, recurrence, evolution/fusion, etc.)
A long-standing tradition of openness

The essence of open source software

- and correlatively of datasets (test suites, parameters, simulations, data models)
  - Note: Zuo et al. showed that GitHub was the most popular repository for hosting COVID-19 datasets

Online proceedings - publication platforms

Inherent understanding of open science issues

- identification, metadata, licencing, citation, versioning, authorship
- e.g. Software Heritage: SoftWare Heritage persistent IDentifiers (SWHIDs), CodeMeta, “Citing software with style” initiative

And of the necessary infrastructures

- from forges to Software Heritage
- high usage of publication repositories such as arXiv or ad hoc repositories (cryptology)
Yearly submission in arXiv (Ginsparg 2021)
Open access to publications in informatics

Inria: 89% of openly available full texts (HAL and arXiv)

According to FoS classes (OpenAIRE), two main related fields:

**Computer & Information Sciences:**
- 98,078 research products
- 65,265 of them (66.5%) being open

**Electrical Engineering, Electronic Engineering & Information Engineering:**
- 2,026,965 research products
- 820,858 of them (40.5%) being open

Specific OA profile (with Mathematics), cf. French open science monitor (next slide):
- favouring green open access (i.e. author’s manuscripts deposited in a publication repository)
- low level of presence in gold open access journals (i.e. with author-pays fees, aka APC - Article Processing Charges)
Opening routes (French publications of 2021, BSO)

<table>
<thead>
<tr>
<th>Hosting type</th>
<th>Open repositories</th>
<th>Publisher &amp; open repositories</th>
<th>Publisher</th>
<th>Open access rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical research</td>
<td>9 %</td>
<td>33 %</td>
<td>19 %</td>
<td>61 %</td>
</tr>
<tr>
<td>(total = 44.4 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fundamental biology</td>
<td>18 %</td>
<td>47 %</td>
<td>10 %</td>
<td>76 %</td>
</tr>
<tr>
<td>(total = 26.1 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earth, ecology, energy and applied biology</td>
<td>23 %</td>
<td>38 %</td>
<td>11 %</td>
<td>73 %</td>
</tr>
<tr>
<td>(total = 15.0 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical sciences and astronomy</td>
<td>34 %</td>
<td>33 %</td>
<td>9 %</td>
<td>75 %</td>
</tr>
<tr>
<td>(total = 14.4 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td>10 %</td>
<td>16 %</td>
<td>29 %</td>
<td>55 %</td>
</tr>
<tr>
<td>(total = 11.7 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>29 %</td>
<td>22 %</td>
<td>12 %</td>
<td>64 %</td>
</tr>
<tr>
<td>(total = 11.0 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Sciences</td>
<td>16 %</td>
<td>17 %</td>
<td>22 %</td>
<td>54 %</td>
</tr>
<tr>
<td>(total = 10.7 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer and information sciences</td>
<td>38 %</td>
<td>18 %</td>
<td>10 %</td>
<td>66 %</td>
</tr>
<tr>
<td>(total = 9.8 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemistry</td>
<td>35 %</td>
<td>28 %</td>
<td>9 %</td>
<td>71 %</td>
</tr>
<tr>
<td>(total = 9.0 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>45 %</td>
<td>27 %</td>
<td>8 %</td>
<td>80 %</td>
</tr>
<tr>
<td>(total = 5.7 k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A certain wealth of scholarly tools, services and data

Generic or thematic infrastructures at the service of the open and free dissemination of their research outputs

- **Software Heritage (foundation in collaboration with Unesco)**
  - 2 billion source files
  - experiment to connect publications and software in HAL
  - central contributor to international efforts within EOSC or RDA

- **Bibliographic databases and digital libraries**
  - DBLP (Schloss Dagstuhl – Leibniz Center for Informatics): since 1993, 6M records of bibliographic references
  - CiteSeeX (Pennsylvania State University): links indexed documents with other sources (DBLP, ACM DL)
  - CEUR (RWTH Aachen University): open-access publication service, 3423 volumes
  - ACM Digital Library: journals and conferences, coupled with the ACM CCS (Computing Classification System)
  - IFIP DL: hosted on HAL, 3 year embargo, 18 000 documents
  - Minimal contribution from IEEE…
Other collections of open bibliographic and citation data

**OpenCitations**: harvests and publishes comprehensive metadata describing academic publications and associated citations

- OpenCitations Index (COCI): citation links
- OpenCitations Meta: bibliographic metadata

**OpenAIRE Graph**: multiple research products, with links to various entities (organisations, funders, funding streams, projects, communities, and data sources)

**CORE**: collection of open access research papers, collecting and indexing research from repositories and journals (300 000 research articles in CS)

**Semantic Scholar**: AI-driven search and discovery tools, open resources
Linking publications with datasets and software

Citations to datasets and software are mostly **informal mentions** in the text body of articles:
- software: only 1-8% of mentions as bibliographic references, 0-0.6% of mentions with PID [2,3]
- ~10% of dataset mentions have PID [4] and datasets are mostly unnamed, e.g.:
  “The **data** has been collected by the UN Comtrade organization, and cleaned by CEPII. ”

**Note:** Data repositories turned out to be limited currently for following and analyzing data usage and creation:
- Metadata debt: lack of affiliation and domain information
- Granularity issues: 1 dataset with 10,000 images gives 10,000 DOI of type “dataset”
- Deposit of datasets in repositories is often not correlated with actual data production
- Necessity to link dataset to real research work to exploit metadata: this requires publications

Publications can be used as **proxies** to the dataset and software usage and creation:
- ensures data is related to research work
- possible to rely on document metadata
Mining full-texts for software mentions

- **Softcite**: software mention detection
  - funding Sloan & Moore Foundations, and French Open Science Plan
  - trained on 4,971 manually annotated documents (37 annotators)
    - [https://github.com/softcite](https://github.com/softcite)

- Automatic characterization of mentions: used / created / shared
  - trained on 3,643 manually annotated sentences
### Mentions to datasets and software

<table>
<thead>
<tr>
<th></th>
<th># documents</th>
<th>share</th>
<th>successful download rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full corpus (2012-2021)</td>
<td>1,426,140</td>
<td>100.00 %</td>
<td></td>
</tr>
<tr>
<td>Full text downloaded</td>
<td>908,567</td>
<td>63.7 %</td>
<td>63.7 %</td>
</tr>
<tr>
<td>→ open access</td>
<td>660,501</td>
<td>46.3%</td>
<td>85.4%</td>
</tr>
<tr>
<td>→ closed access</td>
<td>248,066</td>
<td>17.4%</td>
<td>38.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th># full text documents</th>
<th># mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>processed with Softcite</td>
<td>742,289</td>
<td>3,567,547</td>
</tr>
<tr>
<td>processed with DataStet</td>
<td>621,306</td>
<td>5,607,080</td>
</tr>
</tbody>
</table>

For more information and evaluations, see [Bassinet et al. 2023](#).
Publications mentioning sharing their software

This graph shows, by publication year, the proportion of publications for which a mention of code or software sharing has been detected, among the publications that create code or software. This detection is achieved through an automatic analysis of the full text by the Softcite tool.
Gaps and prospects for further developments

Linking these infrastructures with general research assessment practices

- absence of good linking mechanisms between publications, data and software
  - better cross-referencing mechanisms (software and data citations)
  - automatic detection of data sets and software references in publications
- better coverage of conferences
  - lack of reference authority lists
  - contributions to generic databases (e.g. OpenCitation)
- contribution to the changing assessment landscape (CoARA)
  - Defining CS specific profiles and features
  - Should Informatics Europe take the lead on this?
Dealing with a variety of communities in informatics

From highly theoretical topics

- complexity, graph theory etc.

to practical or experimental areas, often in a multidisciplinary context

- computational biology, natural language processing or digital humanities

Comparing publication profiles across these communities

- types of publishing practices, role of software and data
- providing means for researchers and institutions to generate scholarly dashboards
  - e.g. pattern of international collaborations (authorships in publications, contribution to OSS), cross-disciplinary fertilisation (from/to CS)
Comparison of citing and cited entities in DBLP

Citations from a DBLP publication

23,451,231

31,526,944

54,978,175

Citations to a DBLP publication

25,101,588

56,628,532

Data from COCI sept. dump combine with DBLP october 2021 dump.
Source: https://querty.hypotheses.org/22
Discussion time

What, how, when, who?

- criteria for research assessment in computer science - white paper?
  - existing and prospective criteria
  - from researchers to organisations
- monitoring research (open) production in computer science
  - cf. open science monitor: opening up software reference in individual production and digital collections (institutions, scholarly societies, collections such as CEUR)
  - recommendations concerning software citation (e.g. systematic use of SWHId)
- tooling our community even further to better exploit research productions
  - better reference to conferences (DBLP, CORE, GII-GRIN-SCIE conference rating, etc.)
  - conference maturity level initiative from Informatics Europe
- Contributing to CoARA - coordinated by Informatics Europe?
References


