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

Pierre Paradinas, Laurent Romary

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 - 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

OPEN SCIENCE: JUST SCIENCE DONE RIGHT

Dr. Jon Tennant
@protohedgehog

PUBMET Annual conference
20-21 September, Zadar, Croatia

The specificities of scholarly practices in informatics

Central role of software

- Corresponding assessment practices:
  - *Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria. Computing in Science and Engineering*, Alliez et al. 2019
  - *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

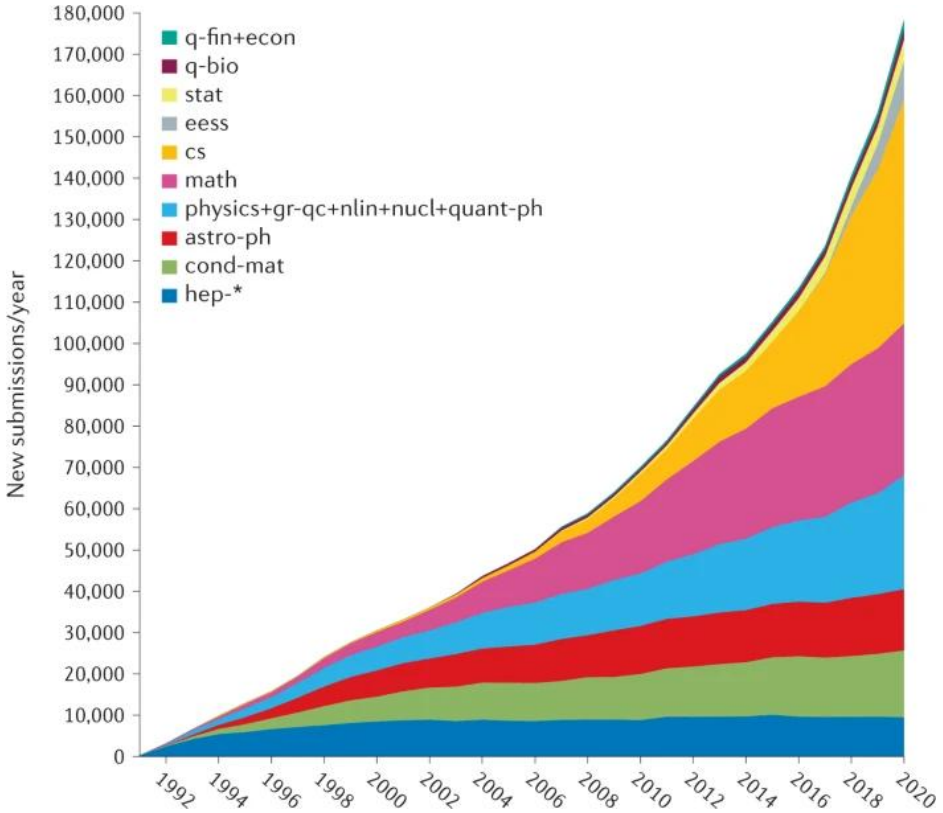
Inherent understanding of open science issues

- 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:

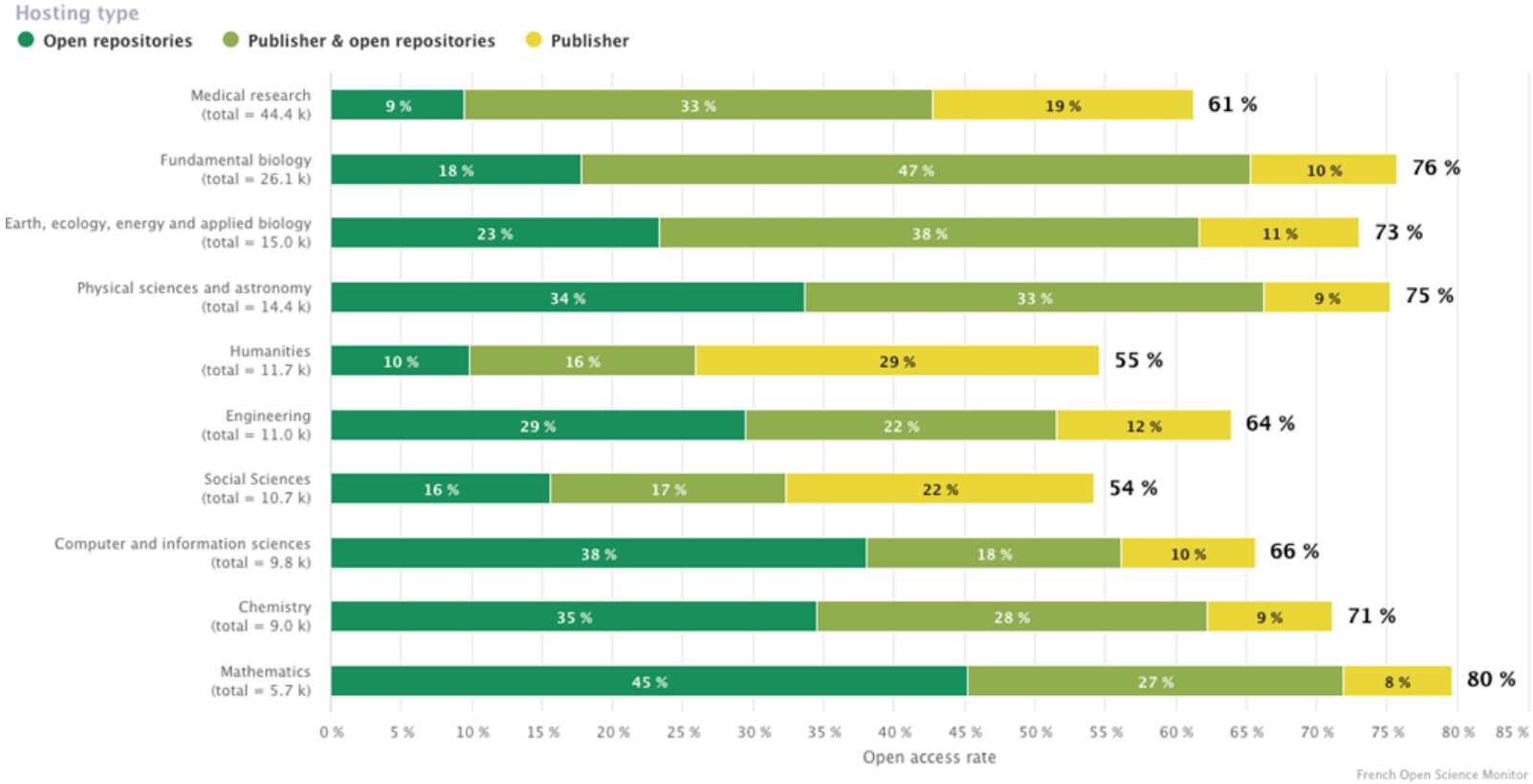
Computer & Information Sciences:
- 98,078 research products
- 65,265 of them (66.5%) being open

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)

https://frenchopensciencemonitor.esr.gouv.fr/

Opening routes (French publications of 2021, BSO)



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:

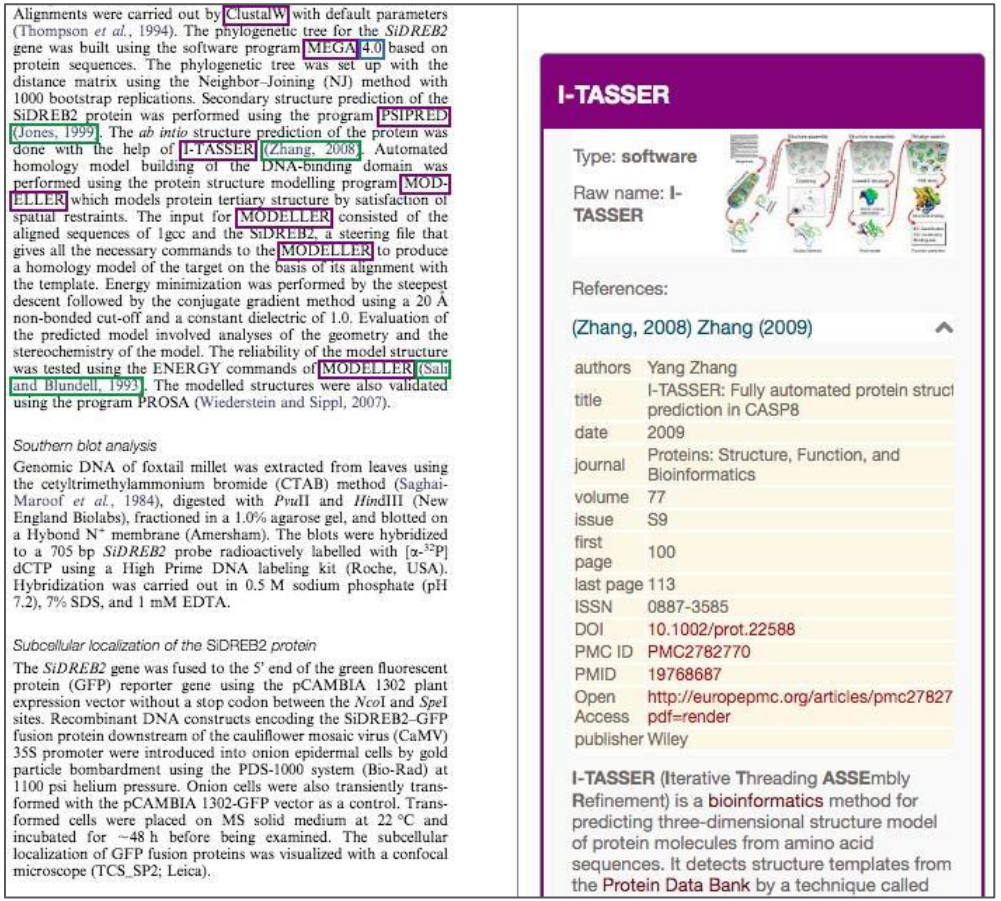
- Metadata debt: lack of affiliation and domain information
- Granularity issues: 1 dataset with 10,000 images gives 10,000 DOI of type "dataset"
- Deposit of datasets in repositories is often not correlated with actual data production
- Necessity to link dataset to real research work to exploit metadata: 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
- Automatic characterization of mentions: **used / created / shared**
  - trained on 3,643 manually annotated sentences



Alignments were carried out by ClustalW with default parameters (Thompson *et al.*, 1994). The phylogenetic tree for the *SiDREB2* gene was built using the software program MEGA 4.0 based on protein sequences. The phylogenetic tree was set up with the distance matrix using the Neighbor–Joining (NJ) method with 1000 bootstrap replications. Secondary structure prediction of the SiDREB2 protein was performed using the program PSIPRED (Jones, 1999). The *ab intio* structure prediction of the protein was done with the help of I-TASSER (Zhang, 2008). Automated homology model building of the DNA-binding domain was performed using the protein structure modelling program MODELLER which models protein tertiary structure by satisfaction of spatial restraints. The input for MODELLER consisted of the aligned sequences of 1gcc and the SiDREB2, a steering file that gives all the necessary commands to the MODELLER to produce a homology model of the target on the basis of its alignment with the template. Energy minimization was performed by the steepest descent followed by the conjugate gradient method using a 20 Å non-bonded cut-off and a constant dielectric of 1.0. Evaluation of the predicted model involved analyses of the geometry and the stereochemistry of the model. The reliability of the model structure was tested using the ENERGY commands of MODELLER (Sali and Blundell, 1993). The modelled structures were also validated using the program PROSA (Wiederstein and Sippl, 2007).

Southern blot analysis

Genomic DNA of foxtail millet was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method (Saghai-Maroof *et al.*, 1984), digested with *Pvu*II and *Hin*dIII (New England Biolabs), fractioned in a 1.0% agarose gel, and blotted on a Hybond N$^+$ membrane (Amersham). The blots were hybridized to a 705 bp *SiDREB2* probe radioactively labelled with [α-^{32}P] dCTP using a High Prime DNA labeling kit (Roche, USA). Hybridization was carried out in 0.5 M sodium phosphate (pH 7.2), 7% SDS, and 1 mM EDTA.

Subcellular localization of the SiDREB2 *protein*

The *SiDREB2* gene was fused to the 5′ end of the green fluorescent protein (GFP) reporter gene using the pCAMBIA 1302 plant expression vector without a stop codon between the *Nco*I and *Spe*I sites. Recombinant DNA constructs encoding the SiDREB2–GFP fusion protein downstream of the cauliflower mosaic virus (CaMV) 35S promoter were introduced into onion epidermal cells by gold particle bombardment using the PDS-1000 system (Bio-Rad) at 1100 psi helium pressure. Onion cells were also transiently transformed with the pCAMBIA 1302-GFP vector as a control. Transformed cells were placed on MS solid medium at 22 °C and incubated for ~48 h before being examined. The subcellular localization of GFP fusion proteins was visualized with a confocal microscope (TCS_SP2; Leica).

I-TASSER

Type: **software**

Raw name: **I-TASSER**

References:

(Zhang, 2008) Zhang (2009)

authors	Yang Zhang
title	I-TASSER: Fully automated protein struct prediction in CASP8
date	2009
journal	Proteins: Structure, Function, and Bioinformatics
volume	77
issue	S9
first page	100
last page	113
ISSN	0887-3585
DOI	10.1002/prot.22588
PMC ID	PMC2782770
PMID	19768687
Open Access	http://europepmc.org/articles/pmc27827 pdf=render
publisher	Wiley

I-TASSER (**I**terative **T**hreading **ASSE**mbly **R**efinement) is a bioinformatics method for predicting three-dimensional structure model of protein molecules from amino acid sequences. It detects structure templates from the Protein Data Bank by a technique called

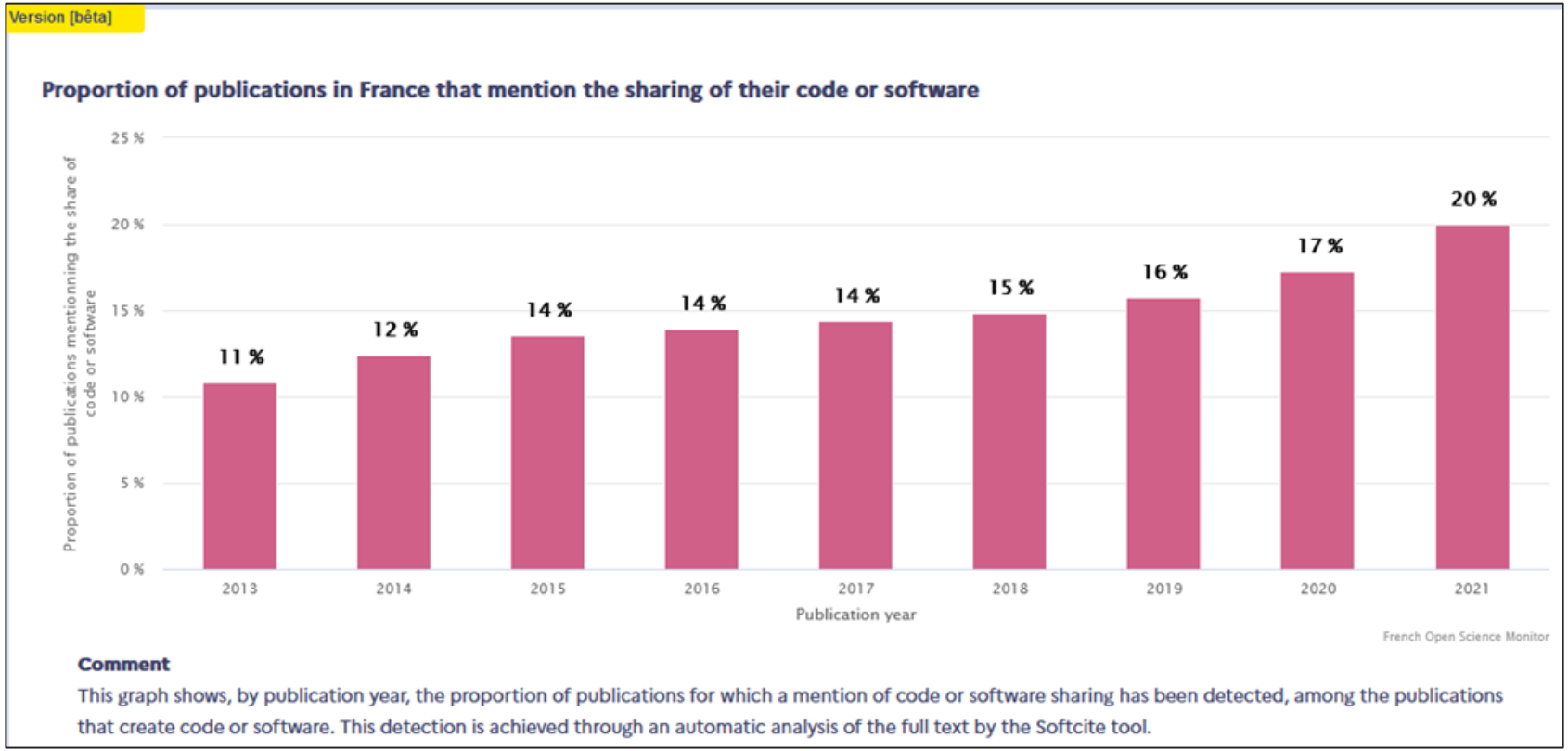
Mentions to datasets and software

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

	# full text documents	# mentions
processed with Softcite	742,289	3,567,547
processed with DataStet	621,306	5,607,080

For more information and evaluations, see Bassinet et al. 2023

Publications mentioning sharing their software



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.

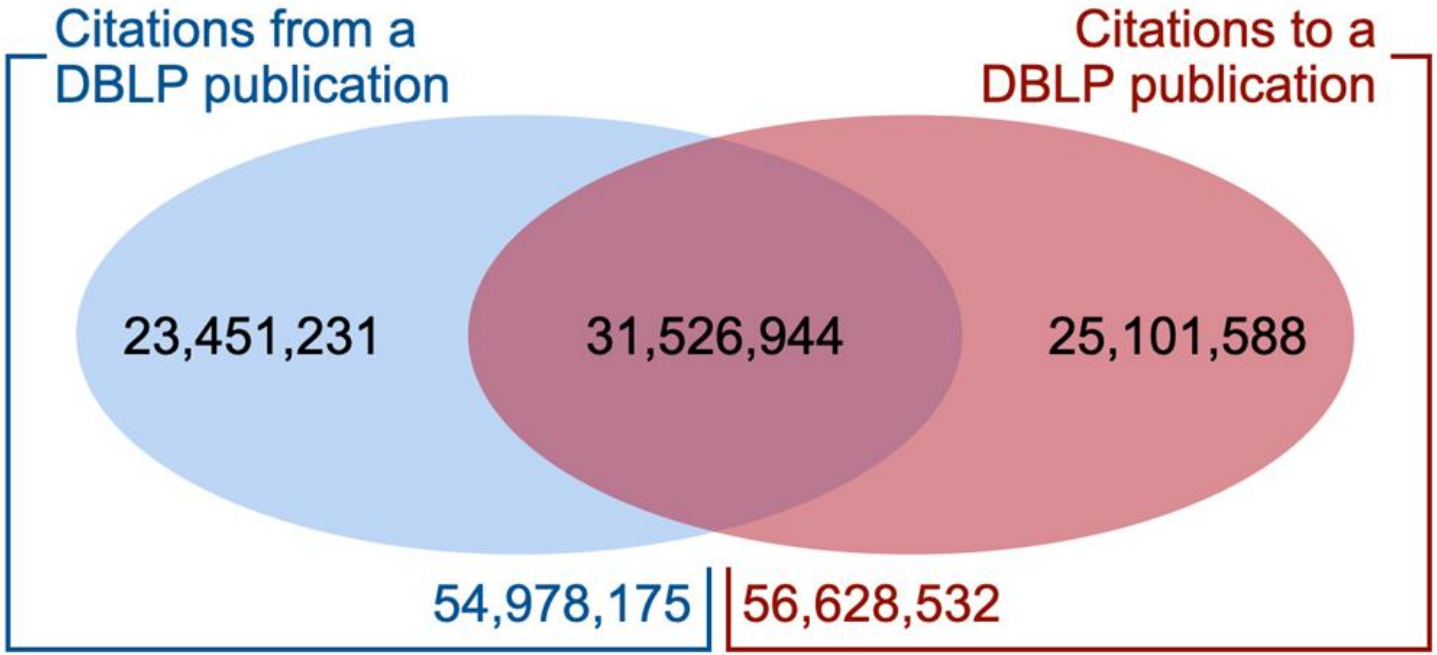
to practical or experimental areas, often in a multidisciplinary context

- 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



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

Alliez P., Di Cosmo R., Guedj B., Girault A., Hacid M.-S., et al.. Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria. Computing in Science and Engineering, 2019, pp.1-14. ⟨10.1109/MCSE.2019.2949413⟩. https://hal.science/hal-02135891

Bassinet A., Bracco L., L'Hôte A., Jeangirard E., Lopez P., et al.. Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France. 2023. https://hal.science/hal-04121339

Bretthauer, D. (2001). Open source software: A history. https://opencommons.uconn.edu/libr_pubs/7/

Canteaut A., M. Angel Fernández, L. Maranget, S. Perin, M. Ricchiuto, et al.. Évaluation des Logiciels. Inria. 2021. https://inria.hal.science/hal-03110723

Ginsparg, P. (2021). Lessons from arXiv's 30 years of information sharing. Nature Reviews Physics, 3(9), 602-603.

Kim, J. Evaluating author name disambiguation for digital libraries: a case of DBLP. Scientometrics 116, 1867–1886 (2018). https://arxiv.org/abs/1806.10540

Kusserow, A., & Groppe, S. (2014). Getting indexed by bibliographic databases in the area of computer science. Open Journal of Web Technologies (OJWT), 1(2), 10-27.

Ley M. (2009). DBLP - Some Lessons Learned. Proc. VLDB Endow. 2(2): 1493-1500. https://dblp.uni-trier.de/xml/docu/dblpxml.pdf

Lin, J., Yu, Y., Zhou, Y., Zhou, Z., & Shi, X. (2020). How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. Scientometrics, 124(1), 555-574.

Vrettas, G., & Sanderson, M. (2014). Conferences vs. journals in computer science. J Assoc Info Sci Technol. https://doi.org/10.1002/asi.23349.

Zuo X, Chen Y, Ohno-Machado L, Xu H. How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles. Brief Bioinform. 2021 Mar 22;22(2):800-811. https://doi.org/10.1093/bib/bbaa331.