How to Open up? (Digital) Libraries at the Service of (Digital) Scholars

Laurent Romary

How to Open up?
(Digital) Libraries at the Service of (Digital) Scholars

Laurent Romary

Inria – team ALMAnaCH

Upside - down

- Libraries as scientific content openers
  - Forgetting the old duties…
- A pragmatic view based on a national and institutional experience
  - Policies and infrastructures
- It’s the digital turn, stupid!
  - Identifying the role and limits of technologies

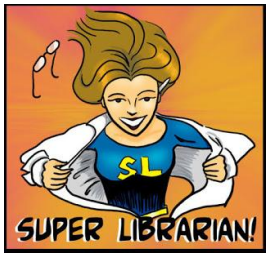
A paradigm change

Provides content

Publishes content

Consumes content

Publishes content



A paradigm change

Once upon a time…

- Collection development
- Budget management
- Cataloguing

… and warehousing

A tough history

- Serial crisis
- Big deals
- Green and gold OA

Diving in the (cold waters of a) new world

- Coordinating scholarly publication management
- Digital content management
  - licencing, persitent identifiers, access statistics
- Drafts, reports, publication, theses, and data

…no management of physical content

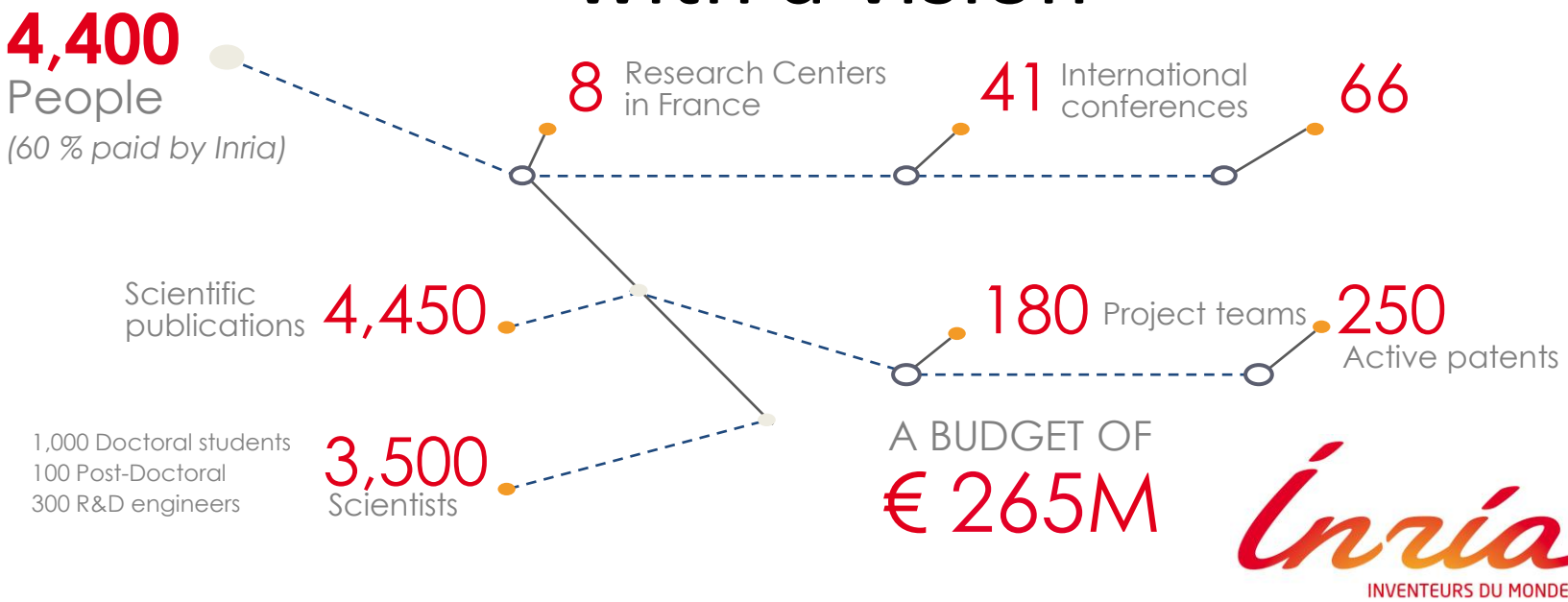
… nor of open access (it's in the genes)

High-Low ➔ low stewardship?

- Research and learning material

A FAVOURABLE CONTEXT

Inria — a research organisation with a vision



Vision for a scientific information policy

- Maximising the dissemination of our scientific assets (visibility and swift dissemination of knowledge), for a reasonable price
- Constitution of a reliable and sovereign institutional corpus (documentation, preservation, access), with clear public governance principles
- Contribution to shaping the scientific communication landscape in terms of editorial processes and usage made of scientific productions

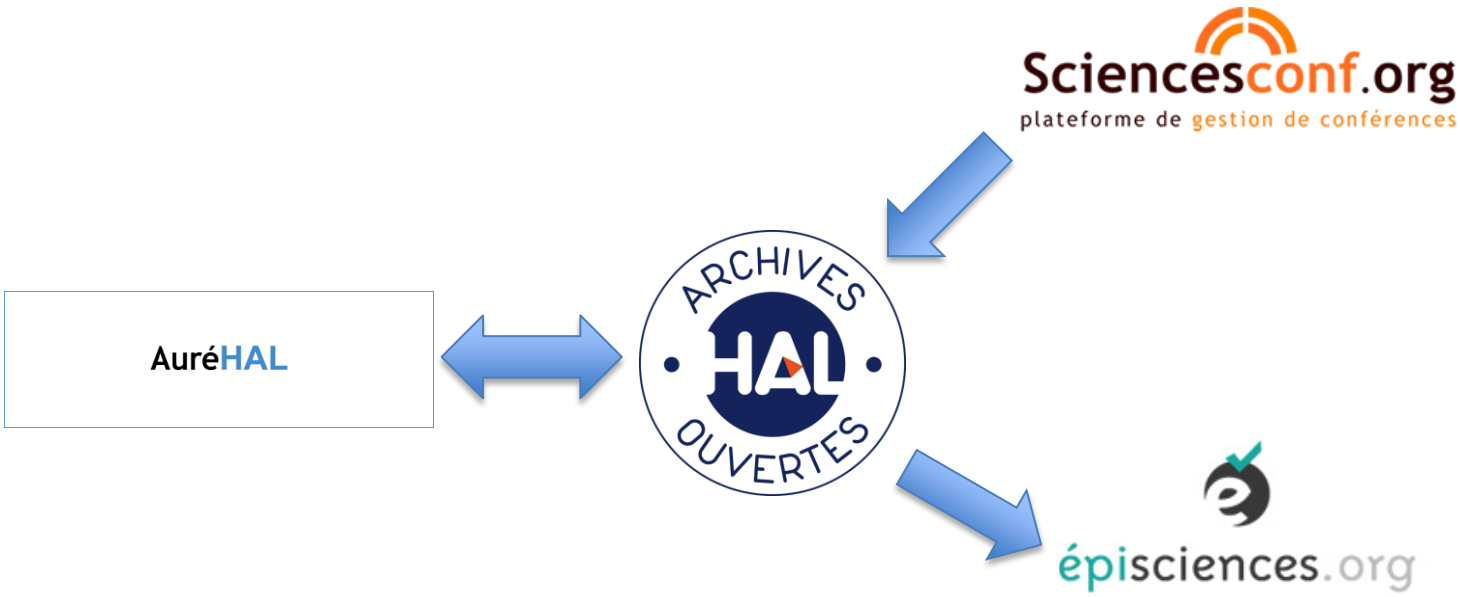
Inria scientific information policy in concrete terms

- Setting up priorities
  - Printed material as disposable goods
  - Deposit mandate on scientific publications
  - Rejecting hybrid open access
  - Engaging in developing new publication models
- Consequences
  - Less collection development on the basis of our acquisitions (national consortia, national licences)
  - More collection development as part of the digital curation activities

Two additional contextual elements

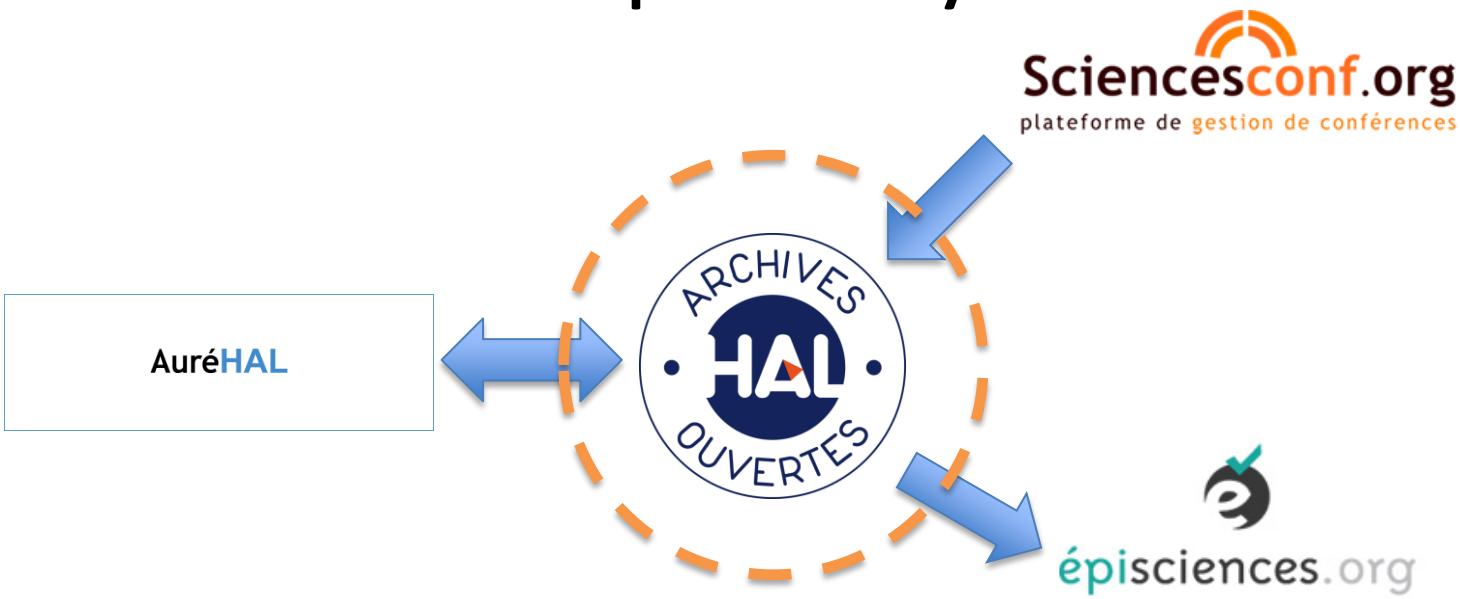
- A strong political support
  - State: *loi pour une république numérique*
    - Published in September 2016
    - Two articles on *open access* and *text and data mining*
  - Higher education and research
    - A wide network of institutions favouring the use of publication repositories
- A technical infrastructure
  - Centred on a national service Unit: CCSD
    - CNRS, Inria, Université de Lyon
  - Development of a comprehensive scientific information management platform

The infrastructure at our service



CCSD

HAL – a multi-purpose publication repository

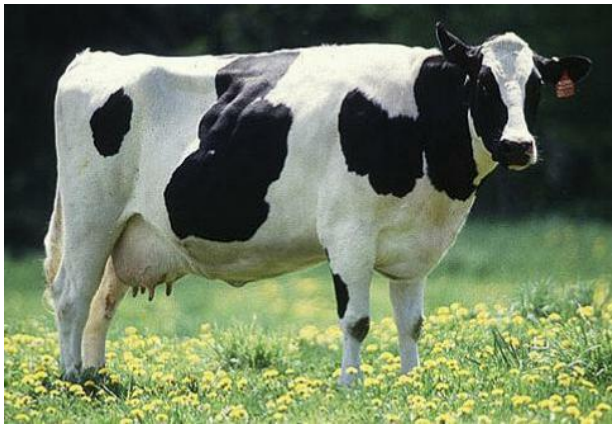


Services	Library support	Technologies

• Centralised publishing environment
• Preprints, articles, theses
• Institutional collections and portal
• Individual webpages
• …
	• Decentralised moderation/curation
	  • Paper, metadata
	• Support to setting up collections (persons, teams, laboratories)
	• Raising awareness: from pre-print to final publications
		• Long term archiving
		• Persistent identifiers
		• Deep meta-data scheme
		• Import-export facilities
		  • TEI, Bibtex, EndNote, etc.
		• And Grobid…

CCSD

Going beyond the limitations of pdf



Cow (structured data)

publishing

text-mining

Hamburger
(unstructured data)

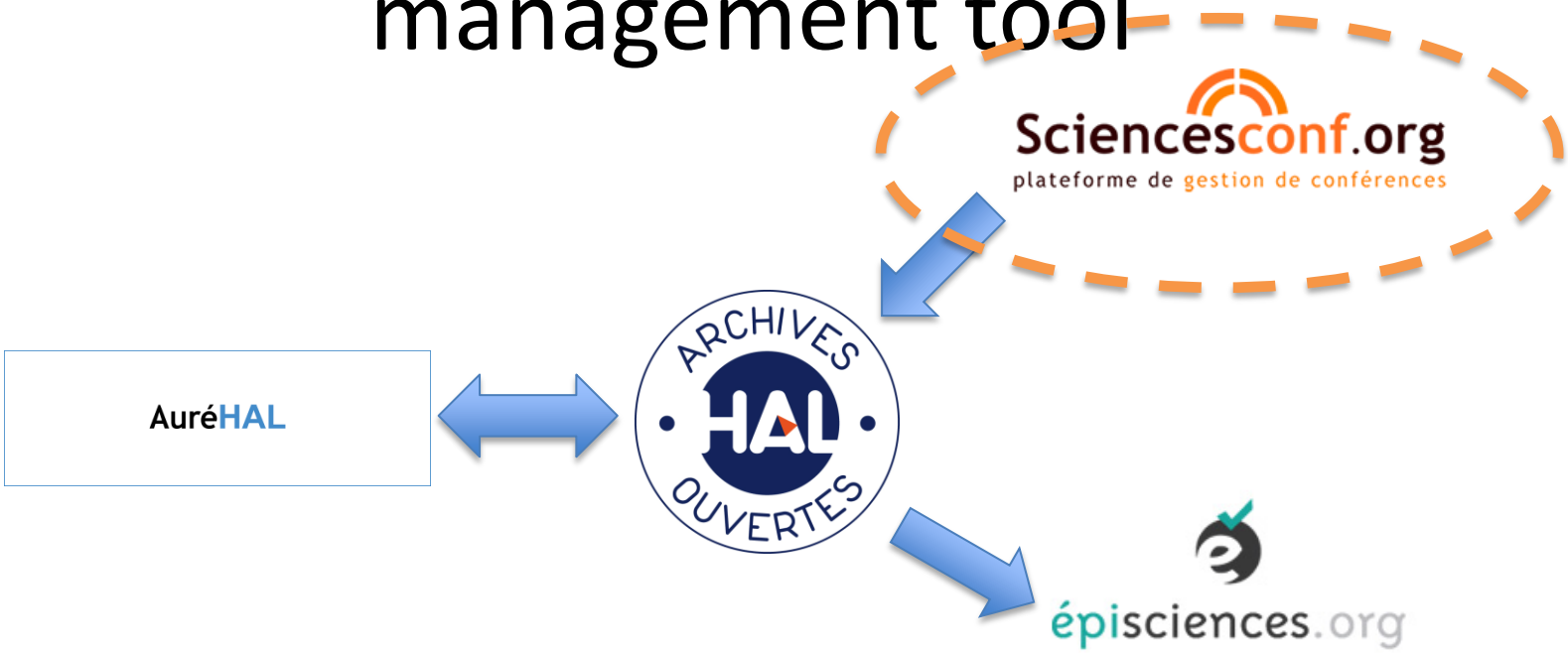
"Converting PDF to XML is a bit like converting hamburgers into cows. You may be best off printing it and then scanning the result through a decent OCR package."
Michael Kay (http://lists.xml.org/archives/xml-dev/200607/msg00509.html)

Inspired from: Duncan Hull

Structuring content

- GROBID: information extraction from PDF documents
  - Meta-data: title, authors, affiliations, abstracts
  - Bibliographical references (with crossref consolidation)
- Standards-based representation
  - TEI (Text Encoding Initiative) as a reference format
- State of the art performance(CRF models)
  - Cf. M. Lipinski, et al., 2013
  - Used at EPO/ResearchGate/Mendeley/CERN/NASA
- Integrated in HAL
  - Automatic meta-data extraction for author's deposit

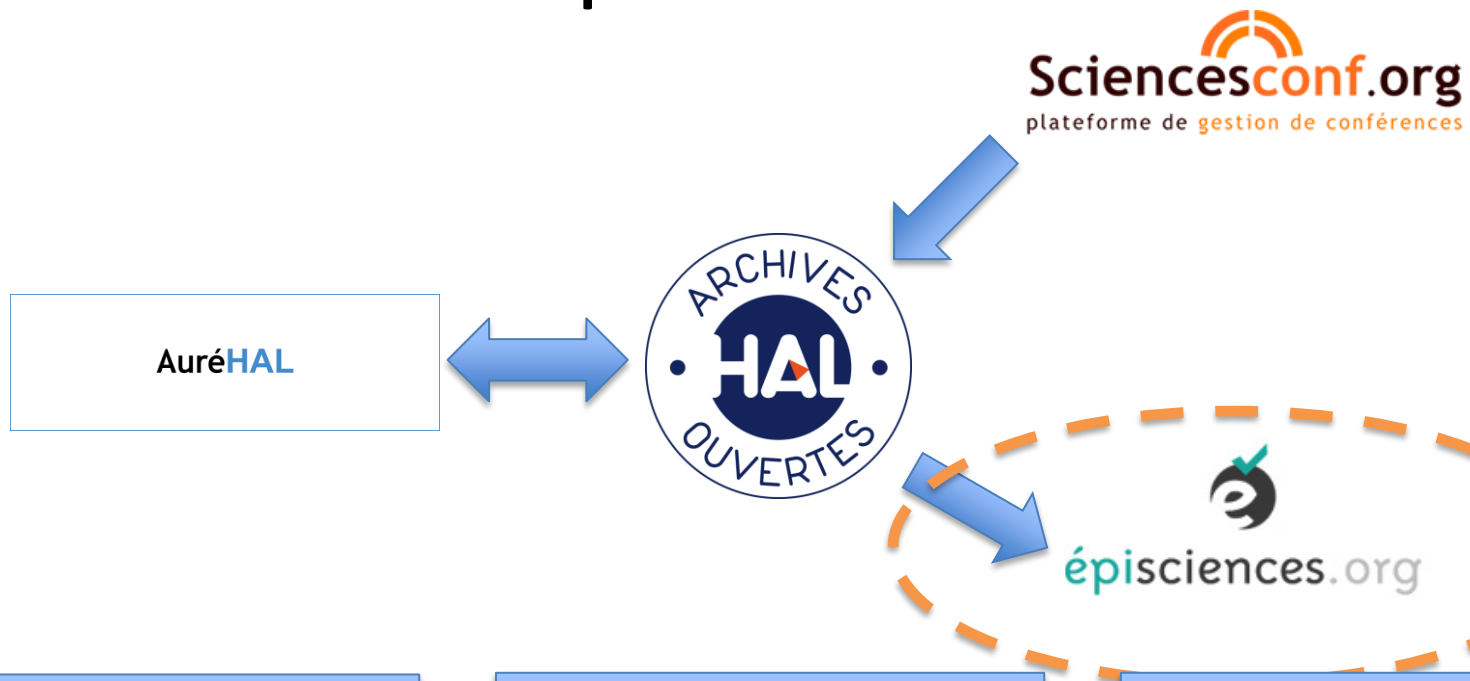
Sciencesconf– a conference management tool



Services	Technologies	Library support
• Online conference management tool • Abstracts, full paper, peer review, conference program, registration	• Modular services • Unique authentication system • Connection to HAL authorities • Upload to HAL	• Supervises the creation of a conference instance • Create conference series • Moderates the integration of the papers as a collection in HAL

CCSD

épisciences – an overlay journal platform

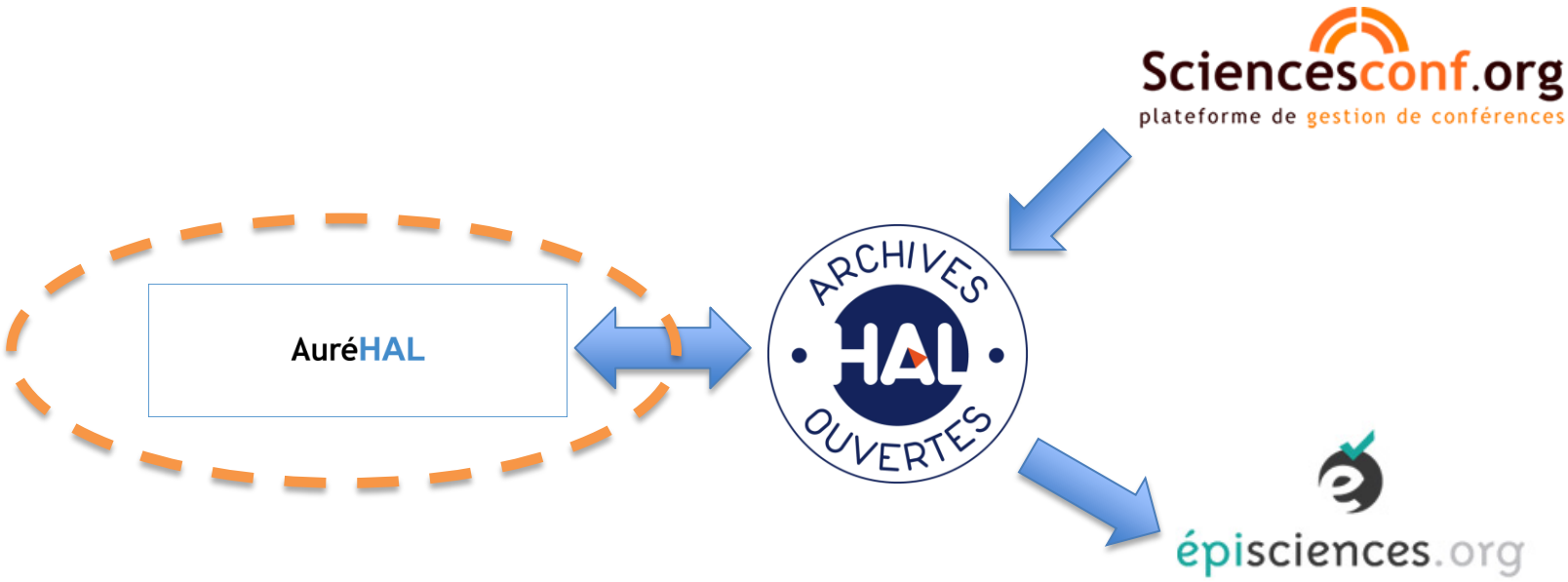


Services	Technologies	Library support

• Overlay certification environment
  • Post publication peer review
  • "Traditional" reviewing environment
• Invited editors - special issues
	• OAI-PMH interface with repositories
	  • Publication: arXiv, HAL, CWI
	  • Data: Nakala, Ortolang
	• Versioning
	• Cataloguing facilities
		• Setting up new journals
		• Editorial support
		• Experimenting new usages: e.g. data journals

CCSD

AureHAL – authorities management



Services	Technologies	Library support

• General purpose authorities
  • Authors
  • Organisations
  • Journals
  • National and EU projects
	• Connection to external databases
	  • ORCID, IdRef, VIAF, etc.
	• Integration in the French organisational framework
	• APIs
	• Standardised export formats (TEI, again)
		• Curates the creation and de-duplication of authorities
		• Confronts with external data bases (e.g. EU projects)

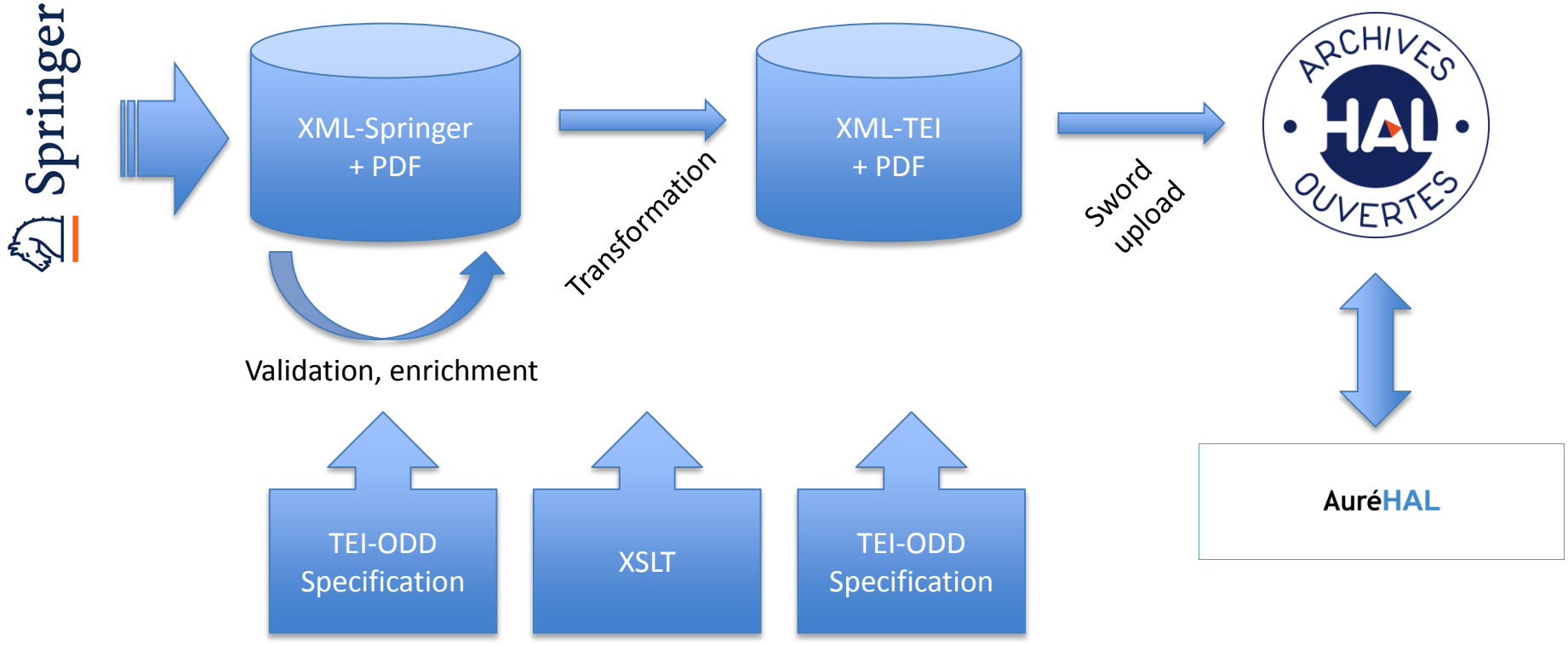
CCSD

AN ENABLING INFRASTRUCTURE FOR ADDITIONAL SERVICES

1. Setting up an institutional digital library

- IFIP - International Federation for Information Processing
  - One of the major scholarly society in Information & Communications
  - Organized in technical committees and working groups
- Objective: setting up a sustainable digital library
  - All volumes publishes as IFIP conferences
  - Visibility, trust, technical facilities
- Publication agreement with Springer
  - Systematic publication in the AICT, LNBIP and LNCS series
  - 3-year embargo before free online dissemination
  - Provision of meta-data (XML Springer) and author post-review manuscripts (PDF)
    - Note: XML Springer formats often contains full-text
- For Inria: an experiment in ingesting and aggregating legacy collections

IFIP DL - architecture



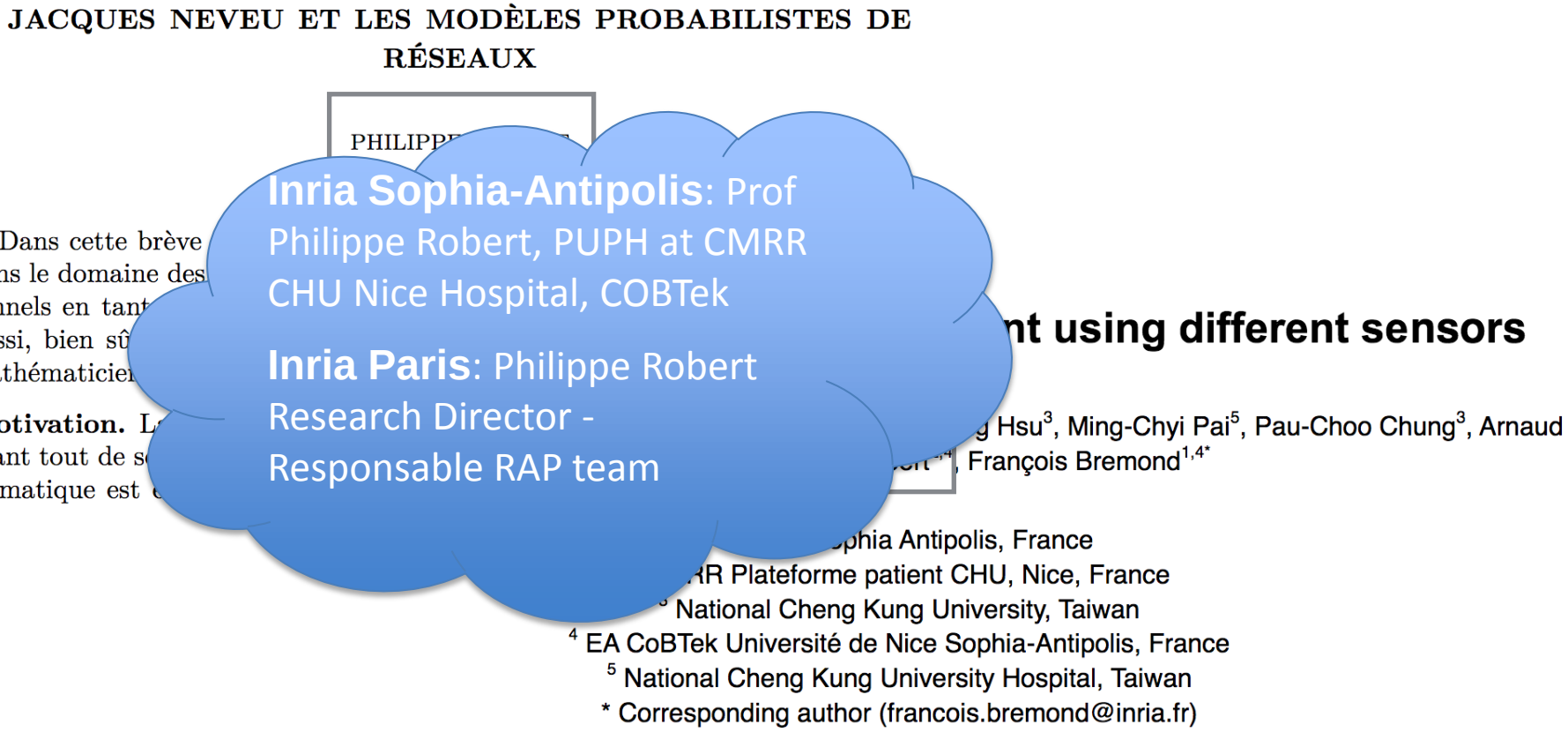
IFIP: technical setting

- Standardized ingestion workflow
- Towards a unified TEI-based format for scholarly papers
  - Cf. EU Peer project (massive ingestion of preprints)
  - Istex, OpenEdition, HAL, EPO
  - More flexible and extensible than JATS (e.g. full text)

IFIP: editorial management

- The central role of library staff
  - Quality check: format, content
  - Collection management: conference series and volumes, technical committees, working groups, etc.
  - Maintenance of authorities in AureHAL: authors, affiliations
    - Approximate affiliations and duplicates
  - Creation and maintenance of XML schemas and XSLT style sheets
- Next step: introducing more automation
  - Objective: limiting boring, repetitive tasks
    - Automatic quality check
    - Entity matching

The entity-matching problem



JACQUES NEVEU ET LES MODÈLES PROBABILISTES DE RÉSEAUX

Dans cette brève
dans le domaine des
sonnels en tant
aussi, bien sû
mathématicien

Motivation. L
avant tout de s
formatique est

Inria Sophia-Antipolis: Prof Philippe Robert, PUPH at CMRR CHU Nice Hospital, COBTek

Inria Paris: Philippe Robert Research Director - Responsable RAP team

nt using different sensors

g Hsu[3], Ming-Chyi Pai[5], Pau-Choo Chung[3], Arnaud
, François Bremond[1,4*]

ophia Antipolis, France
RR Plateforme patient CHU, Nice, France
[3] National Cheng Kung University, Taiwan
[4] EA CoBTek Université de Nice Sophia-Antipolis, France
[5] National Cheng Kung University Hospital, Taiwan
* Corresponding author (francois.bremond@inria.fr)

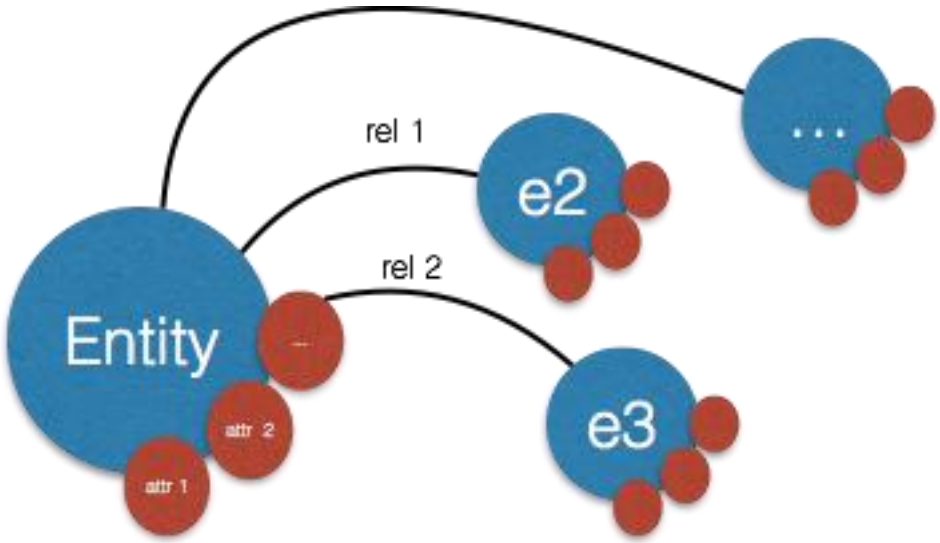
Purpose: Older people population is expected to grow dramatically over the next 20 years (including Alzheimer's patients), while the number of people able to provide care will decrease. We present the development of medical and information and communication technologies to support the diagnosis and evaluation of dementia progress in early stage Alzheimer disease (AD) patients. **Method**: We compare video and accelerometers activity assessment for the estimation of older people performance in instrumental activities of daily living (IADL) and physical tests in

Mapping

The mapping is a set of domain specific wrappers transforming the input/output from/to the internal data model.

The internal data model is composed of:

- entities
- attributes
- relations



HAL authors

Local attributes: surname, last name, title, email, etc

Comparison using distances depending upon the type of the attribute:

e. g. Person Name is computed using Dice Sorensen, Cosine similarity and Jaro Winkler.

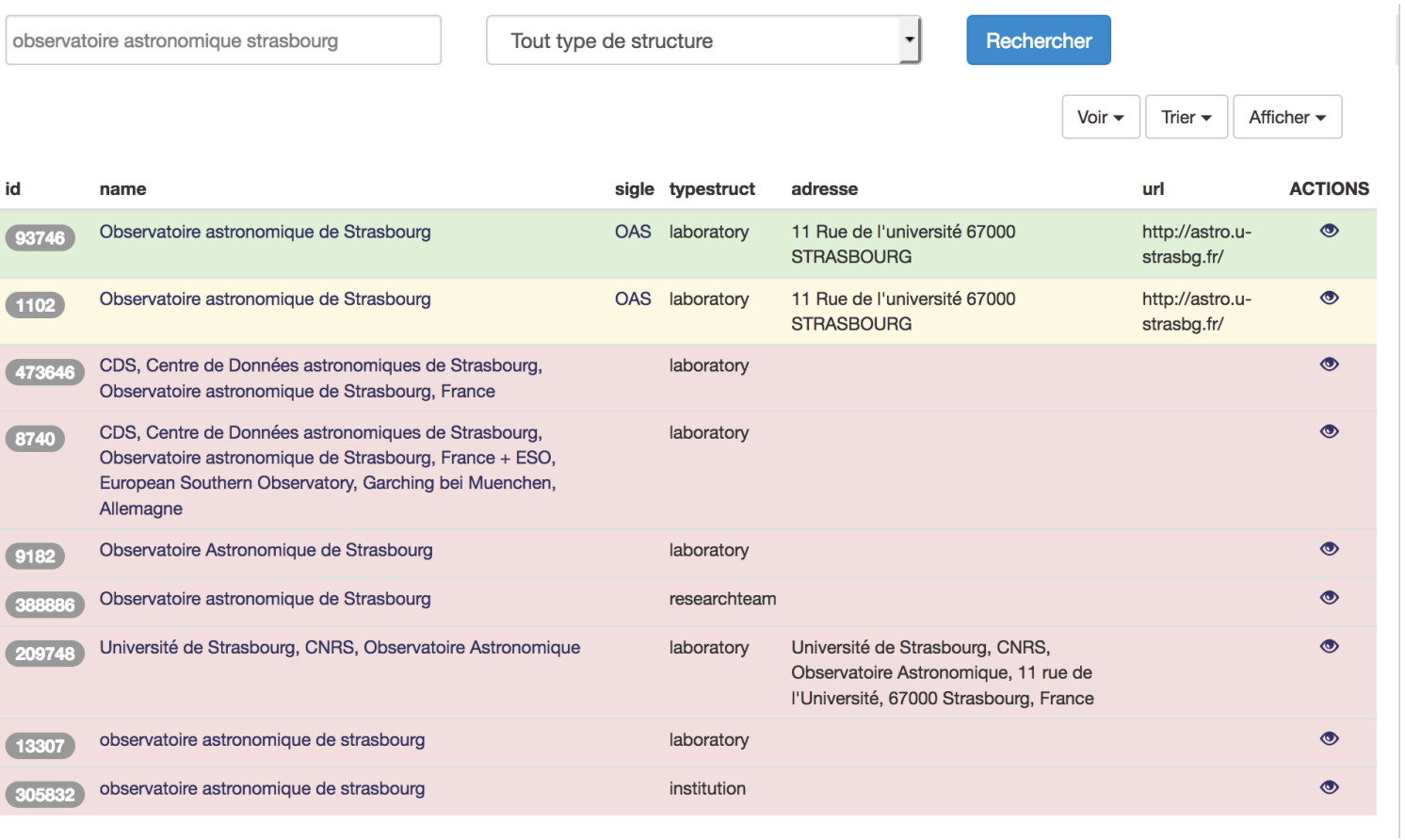
Relation attributes: co-authorship, affiliations, popularity*, years of activity*, etc.

Relations are computed using affiliation information

Preliminary results are promising, F-1 score of 0.964

HAL organisation

Representation and errors are more frequent



id	name	sigle	typestruct	adresse	url	ACTIONS
93746	Observatoire astronomique de Strasbourg	OAS	laboratory	11 Rue de l'université 67000 STRASBOURG	http://astro.u-strasbg.fr/	
1102	Observatoire astronomique de Strasbourg	OAS	laboratory	11 Rue de l'université 67000 STRASBOURG	http://astro.u-strasbg.fr/	
473646	CDS, Centre de Données astronomiques de Strasbourg, Observatoire astronomique de Strasbourg, France		laboratory			
8740	CDS, Centre de Données astronomiques de Strasbourg, Observatoire astronomique de Strasbourg, France + ESO, European Southern Observatory, Garching bei Muenchen, Allemagne		laboratory			
9182	Observatoire Astronomique de Strasbourg		laboratory			
388886	Observatoire astronomique de Strasbourg		researchteam			
209748	Université de Strasbourg, CNRS, Observatoire Astronomique		laboratory	Université de Strasbourg, CNRS, Observatoire Astronomique, 11 rue de l'Université, 67000 Strasbourg, France		
13307	observatoire astronomique de strasbourg		laboratory			
305832	observatoire astronomique de strasbourg		institution			

HAL organisation

Local attributes: name, address, region, country, etc

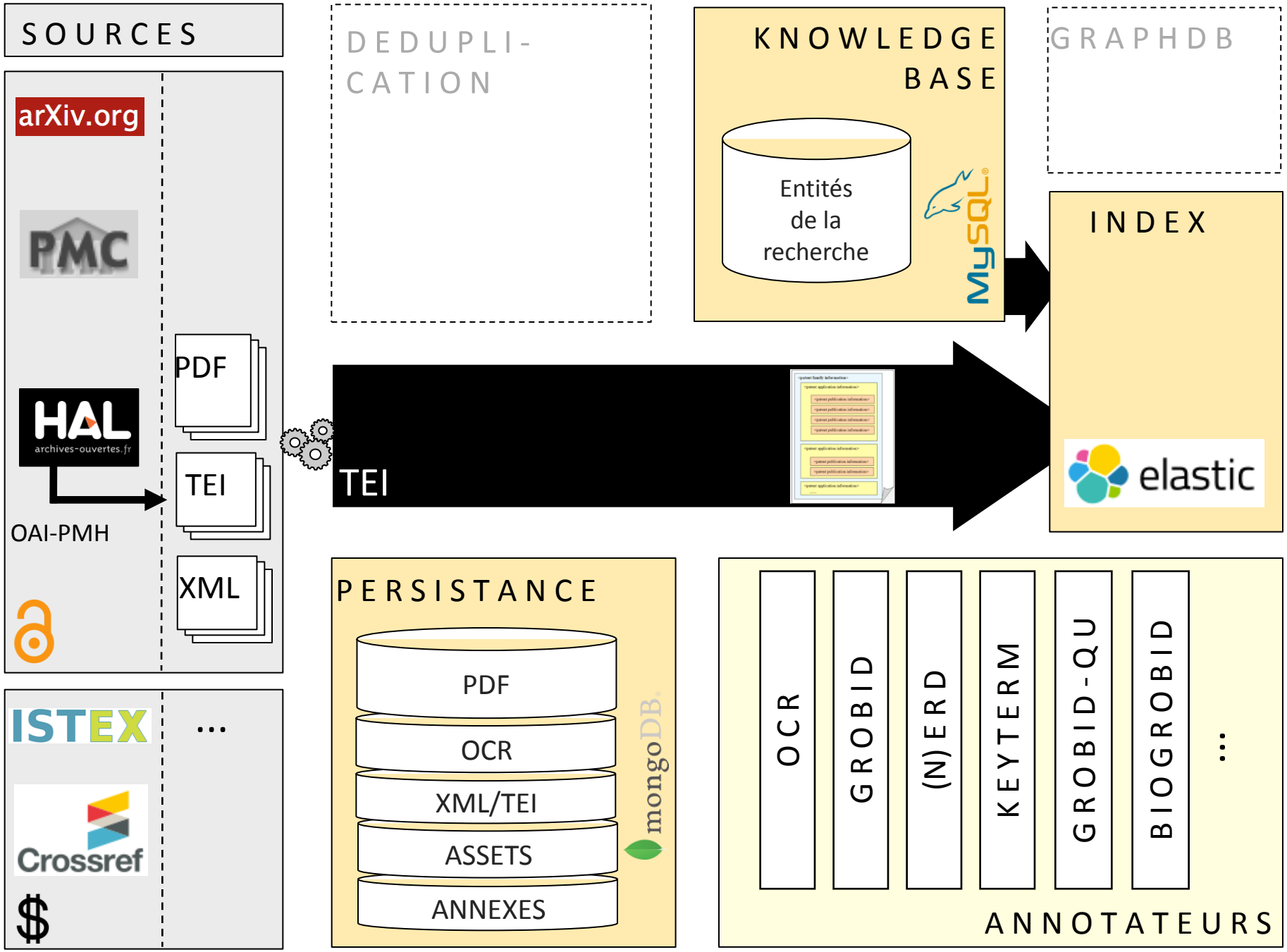
Relation attributes: mother-child (e.g Inria -> Inria Paris), authors affiliated, global relations*

Like for authors, relations are computed using complementary information.

Preliminary results are still too low, F-1 score of 0.345

2. Exploiting repository contents: AnHALytics

- Objectives
  - Designing a scholarly dashboard
    - Scientific profiles: publications and authors
  - Experimenting various data extraction mecanisms
- Reference content from the HAL publication repositories
  - Meta-data: title, periods, authors, affiliations
  - Full-text: full-text indexing, conceptual search
- Technical background
  - NERD and Grobid full text
- Librarians are shaping the service with developpers
  - Researchers, Inria management


SOURCES
arXiv.org
PMC
HAL
archives-ouvertes.fr
OAI-PMH
PDF
TEI
XML
ISTEX
...
Crossref
$
DEDUPLI-
CATION
KNOWLEDGE
BASE
Entités
de la
recherche
MySQL
GRAPHDB
INDEX
elastic
TEI
PERSISTANCE
PDF
OCR
XML/TEI
ASSETS
ANNEXES
mongoDB.
OCR
GROBID
(N)ERD
KEYTERM
GROBID-QU
BIOGROBID
...
ANNOTATEURS

Making things concrete

- AnHALytics - demo Inria

Plugging additional technologies - 1

- Grobid quantities
  - Identifying measured quantities in scientific documents
  - Value, unit, measured entity
  - Multiple search scenarios in a scholarly context

Example (example 2, streptomycin)

The cells were washed three times with RPMI1640 medium (Nissui Pharmaceutical Co.). The cells (1 x107) were incubated in RPMI-1640 medium containing 10% calf fetal serum (Gibco Co.), 50 µg/ml streptomycin, 50 IU/ml of penicillin, 2-mercaptoethanol (5 x 10-5 M), sheep red blood cells (5 x 106 cells) and a test compound dissolved in dimethyl sulfoxide supplied on a microculture plate (NUNC Co., 24 wells) in a carbon dioxide gas incubator (TABAI ESPEC CORP) at 37°C for 5 days.

A solution of 1.18 g (4.00 mmols) of the Compound a obtained in Reference Example 1, 0.39 g (4.13 mmols) of 4-aminopyridine and 20 ml of toluene was heated to reflux for 2 hours. After cooling, the reaction mixture was poured into 1 N sodium hydroxide aqueous solution, and washed twice with chloroform. 2 N Hydrochloric acid aqueous solution was added to the aqueous layer and the precipitated white crystals were filtered and dried to give 0.73 g (yield: 53%) of Compound 3.

Atomic value

raw value: **50**

raw unit name: **µg/ml**

normalized value: **0.00005**

normalized unit name: **kg/m^4**

quantified (experimental):

raw: **streptomycin**

normalized: **streptomycin**

Plugging additional services - 2

- Re-publishing content
  - Grobid-TEI as a pivot format in the publishing environment
  - Generation of multiple derived format
    - HTML
    - ePub
    - Braille
- How far are we from this?
  - Improving the performances of Grobid (e.g. book)
  - Partially implemented in various initiatives
    - Revues.org
    - Istex
- There again, a central role for the library staff
  - Identifying user expectations
  - Maintaining formats and transforms

Conclusion

- Digital sovereignty
  - Mastering scholarly content at all stages of the publication process
  - Mastering the whole scholarly process
    - Understanding where and when we need to resort to the private sector
  - Mastering interoperability by means of open formats
  - A general culture of openness
- Focusing our efforts
  - More brain, budget etc. in shaping this digital turn
  - Stop loosing time in “negotiating” open access: let’s just do it
- Dedicated research activity on scholarly information management
  - TDM breakthroughs at the service of scientific information
  - Dealing with the quickly evolving digital context
  - Coupling service units with adequate research teams
- Training future digital librarians
  - Formats, standards and related technologies
  - Data mining and visualisation

References

- Mabe M. A., Scholarly communication: A long view, *New Review of Academic Librarianship*, 2010 - Taylor & Francis
- Hopfield J. J. , Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. U.S.A.*, 79 (1982), pp. 2554–2558
- Berthaud Christine , Laurent Capelli, Jens Gustedt, Claude Kirchner, Kevin Loiseau, et al.. EPISCIENCES - an overlay publication platform. *ELPUB2014*, Jun 2014, Thessalonique, Greece. <http://www.ebooks.iospress.nl/publication/36552>. <10.3233/978-1-61499-409-1-78>. <hal-01002815v2>
- Romary Laurent. Scholarly Communication. Mehler, Alexander and Romary, Laurent. *Handbook of Technical Communication*, de Gruyter, 2012, 978-3-11-022494-8. <inria-00593677>