Digital Humanities, Open Science & Infrastructures... and beyond

14:00-14h25 Laurent Romary – Inria and BBAW – Open science policies and infrastructures: fostering cross-border collaborations

14:25-14:50 Prof. Dr. Andreas Witt – Leibniz Institute for the German Language – Text+: An Infrastructure for Language Data within the German NFDI

14:50-15:15 Alexander Geyken and Axel Herold – Berlin-Brandenburg Academy of Sciences and Humanities- Lexical resources in NFDI/Text+

15:15-15:40 Georg Rehm – DFKI and HU Berlin – Selected Infrastructure Activities around Language Data at DFKI

15:40-16:05 Toma Tasovac – President of the DARIAH Board of Directors- Why Open Science Needs Research Infrastructures and Vice Versa: The Case of DARIAH

16:05-16:40 Stephan Grumbach and Sébastien Grappe– Inria and Centre Marc Bloch – Measuring futurability

Wrap-up

OPEN SCIENCE POLICIES AND INFRASTRUCTURES: FOSTERING CROSS-BORDER COLLABORATIONS

Laurent Romary
Director for Scientific Information and Culture

Open science and infrastructures

1. Why thinking of open science at all?

Serving science and its actors

2. The situation in France and at Inria

Our infrastructures

3. Room for cross-border collaborations

Open science at the service of infrastructures

Inria

Scientific information and open science — why do we need national and institutional policies?

Supporting researchers in managing and communicating scientific research productions

- Publications, data, software, methods
- At the service of science and mankind

Impact

- Visibility, citability, measure

Scientific transparency

- Proving, comparing, reproducing

Quality

- Formats, metadata, authorities
- Researchers/library specialists

Budget control

- Subscriptions, APCs, investments

Ensuring our (digital) sovereignty

- Setting up a corpus of our research productions
- Sustainable public infrastructures, text and data mining, indicators

Cf. presentation by Olivier Colliot this morning on reproducibility…

Open science: (French) national context

Deuxième Plan national pour la science ouverte

GÉNÉRALISER LA SCIENCE OUVERTE EN FRANCE 2021-2024

2016: Loi pour une république numérique (Lemaire)

Allows self archiving with maximal embargo periods for accepted authors' manuscripts

Open data became mandatory for all publicly funded data, including research data

2018: Premier plan national pour la science ouverte

Creation of French Open Science Committee

Setting up a National Open Science Fund

2021: Deuxième plan national pour la science ouverte

New action plan with 70+ items

The pillars of open science: publications, research data, research source codes and software, transforming scholarly practices

https://www.ouvrirlascience.fr/second-national-plan-for-open-science/

Main infrastructures for open science

HAL – the multidisciplinary national publication repository

Publications, theses, reports etc. (1 million+ documents, 3 million+ references)

Additional services: conference management, overlay journal platform (Episciences)

Inria is one of the hosting institutions of HAL (with CNRS and INRAe)

Software Heritage – the international repository that collects, preserves, and shares all software that is publicly available in source code form

Harvests all (GitHub, Gitlab etc.) open software forges - ~20 billion source files

Collaboration with Unesco

Inria is a co-founder to SH – e.g. linking HAL and Software Heritage

Recherche Data Gouv – the national open research data repository

Initiated by the French national committee for open science

Dataverse solution - set up by INRAe

Officially on air since 8 July 2022

Today's Software Heritage Figures

Do we already have your code?

We harvest publicly available source code from many software projects and keep up with development happening there. As of today our archive already contains and keeps safe for you:

Source files	Commits	Projects
19,041,725,765	4,128,578,384	300,048,166

Inria scientific information policy in concrete terms

Deposit mandate on all scientific publications (in HAL, CC-BY)

- **Condition to appear in annual research reports**

- Encouragement towards preprints

Central budget for APCs (aka SIS - Supervised injection sites)

- Forbidding hybrid open access

- **Publisher's PDF to be deposited in HAL (with equal licence)**

- Management of a national dashboard of costs and journals (OpenAPC)

- No refunding requests to funders (FR, EU)

Engaging in developing new publication models and infrastructure

- **Editorial support to *Episciences* based journals**

- Investment and support to *Software Heritage*

- **Editorialization (IFIP DL, conferences)**

Research data support group

- Data management plan, support to the use of the new Recherche Data Gouv infrastructure

- National network of scientific contacts/champions for research data

- **New**: preventing data set deposit on journal platforms

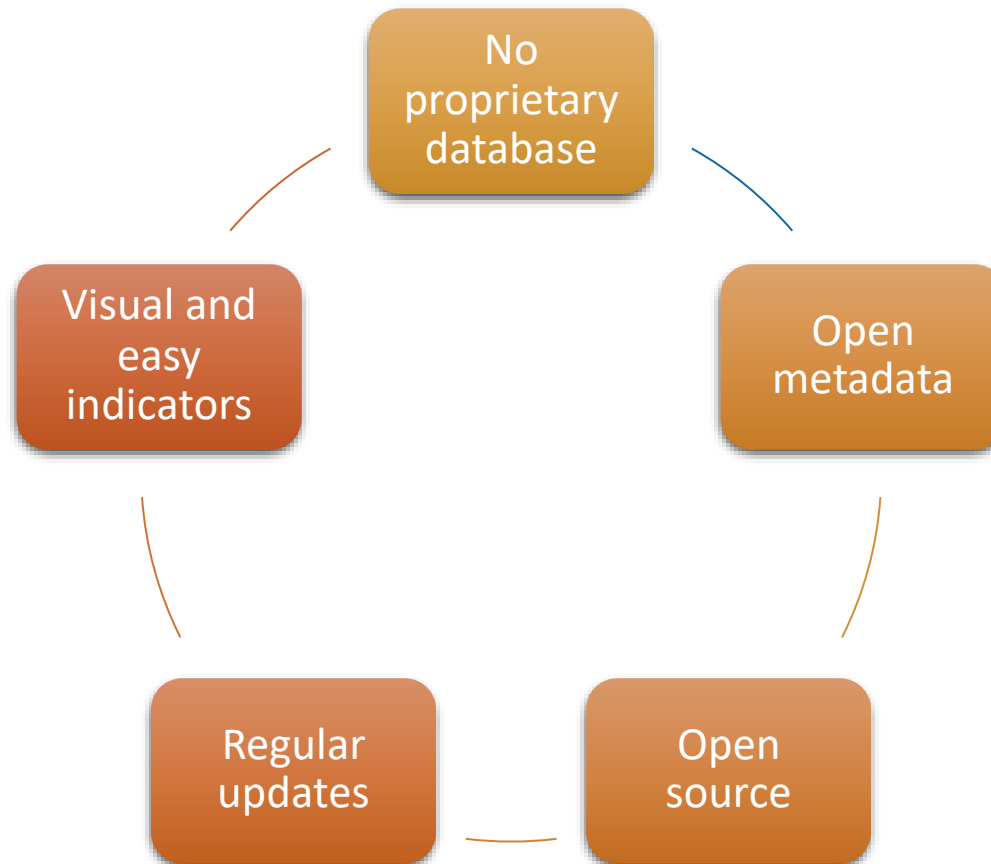**Printed material as disposable goods**

- Creation of a central collection of reference works

- **On going digitization of legacy publications (on HAL)**

Engaging in text and data mining

- Contribution to the national open science monitor
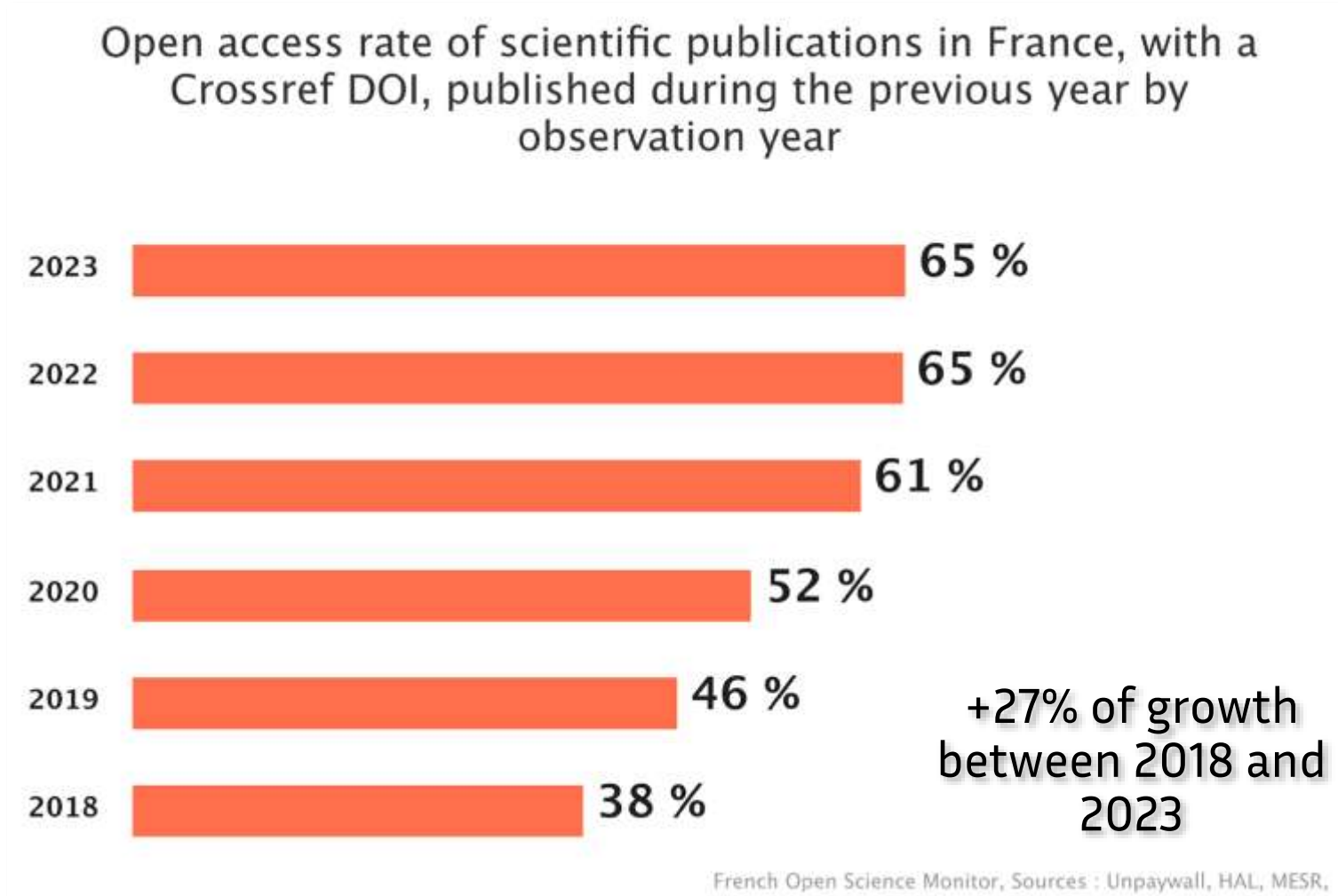
- Software and data set citations

Monitoring open science: the French open science monitor (BSO – baromètre de la science ouverte)

- A need for a national open science monitor with open indicators
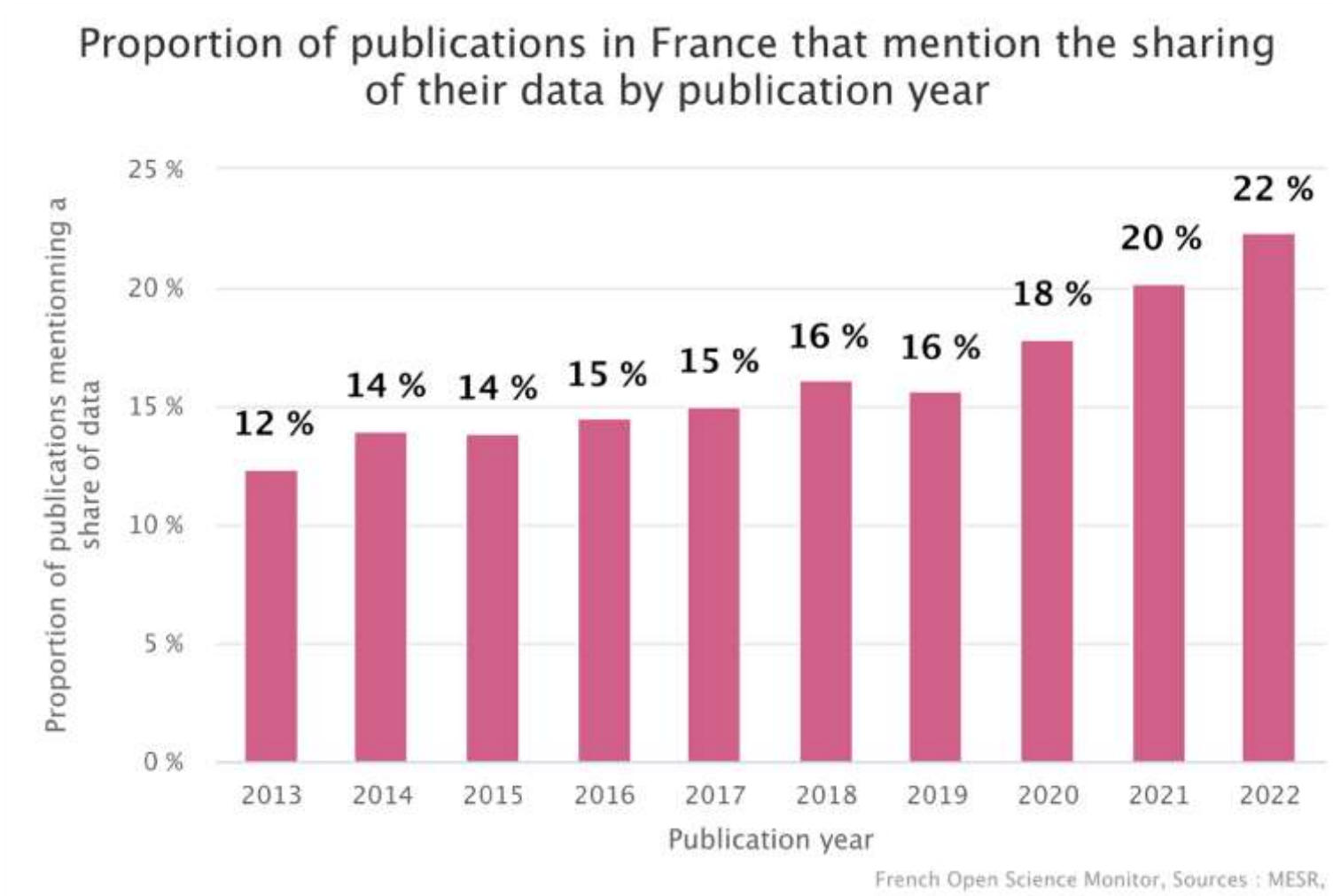
- What were the requirements?

No proprietary database

Open metadata

Open source

Regular updates

Visual and easy indicators

The results of the latest release (2024): results

Open access rate of scientific publications in France, with a Crossref DOI, published during the previous year by observation year

Year	Rate
2023	65 %
2022	65 %
2021	61 %
2020	52 %
2019	46 %
2018	38 %

+27% of growth between 2018 and 2023

French Open Science Monitor, Sources : Unpaywall, HAL, MESR,

Inria

The results of the latest release (2024): datasets

Proportion of publications in France that mention the sharing of their data by publication year

The results of the latest release (2024): software

Proportion of publications in France that mention the sharing of their code or software by publication year

French Open Science Monitor, Sources : MESR,
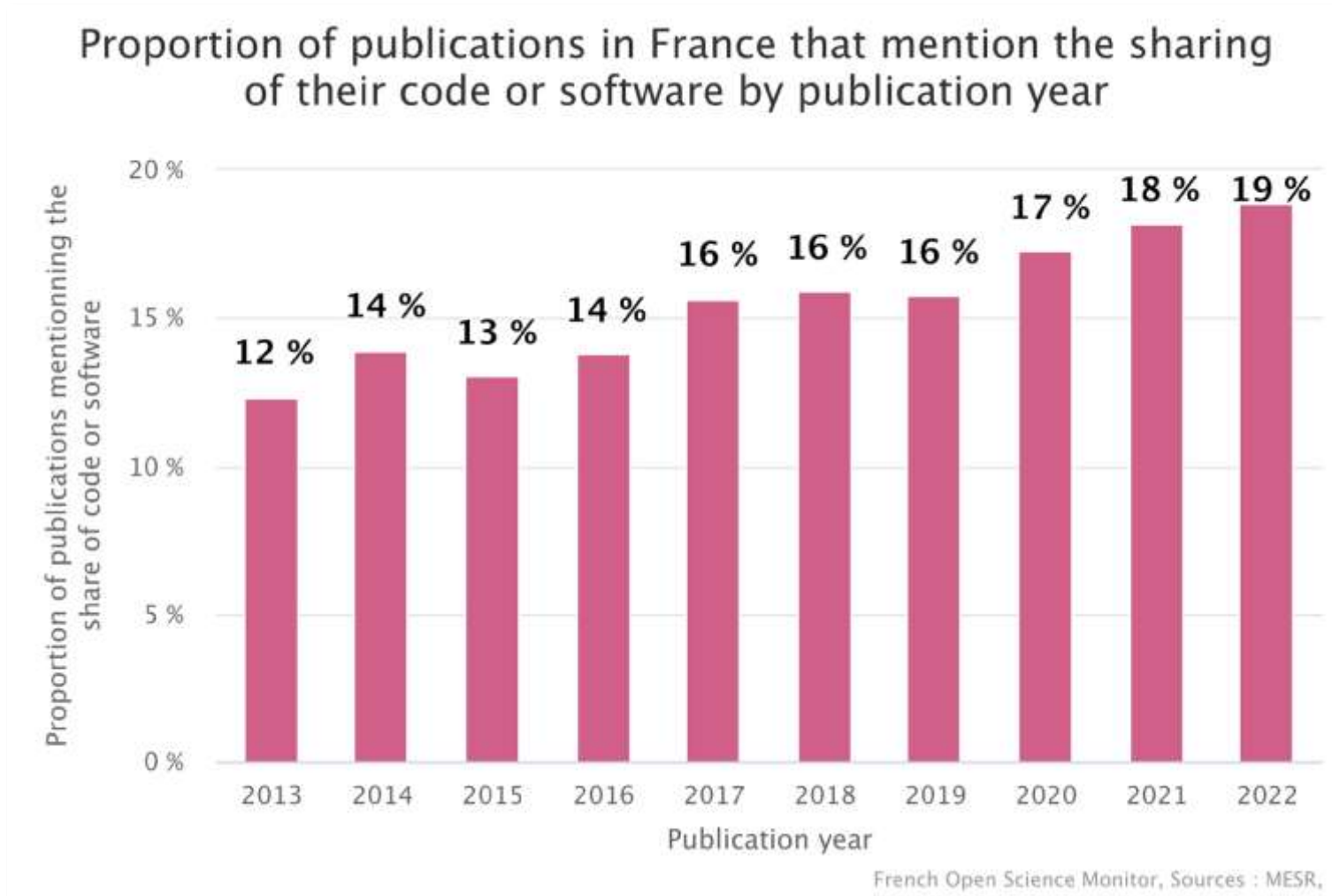
What we acheived at Inria...
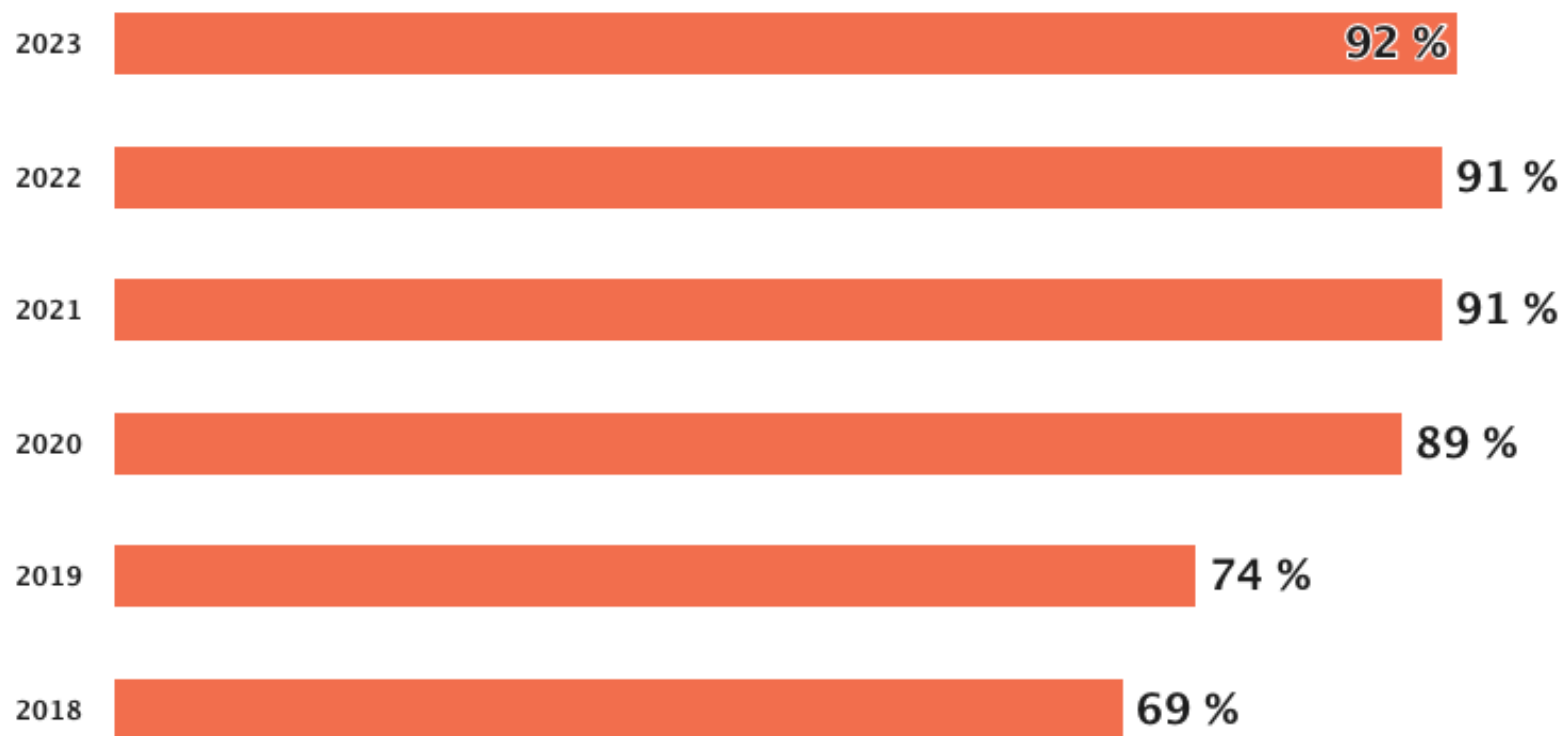
Baromètre français de la Science Ouverte – CC-BY MESR

Epilogue: back to infrastructures

National and European infrastructures need strong open science policies:

- More visibility
  - Citation, user base
- More transparency
  - Monitoring, accountability
- More innovation
  - Letting more communities experiment

And there is work to be done:

- Dissemination: publication, data, software
- Standardisation: testing, contributing
- Education: shaping future research practices
- New models: technical, editorial, economical adapted to the served communities
- Looking ahead...

Merci.

Misc.

Interoperability standards

"I am sending you a document compliant with ISO 8601"

"2022-07-04"

"Received loud and clear. I'll check the standard to understand."

"Oh, you mean the 4th of July of year 2022 in the Gregorian calendar!"

Exchange of data, pooling data from various origins, interoperability between software components, comparability of results => essential for Artificial Intelligence components
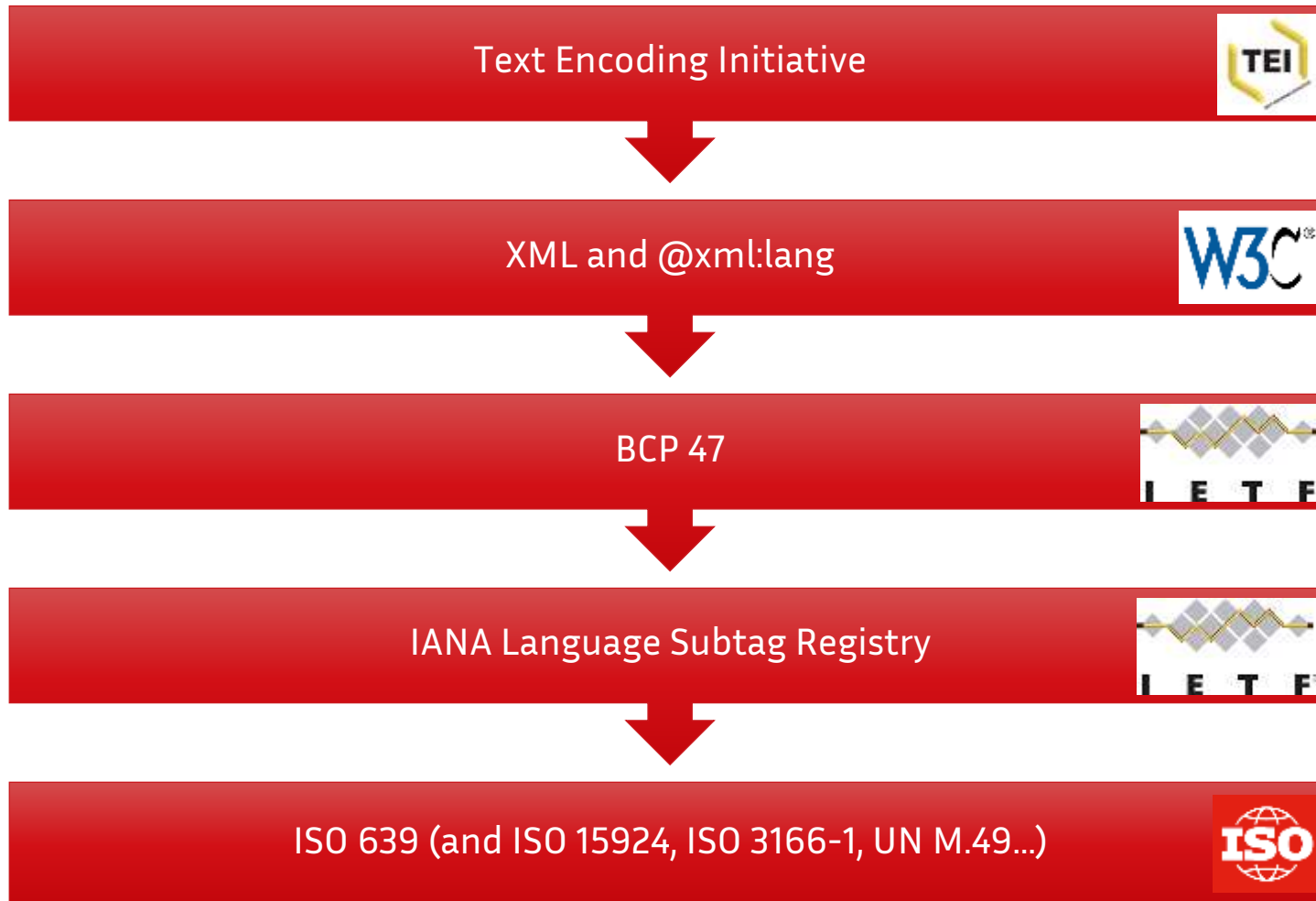
How many standards do we need?

Combining standards is standard practice

```
<quote xml:lang="fr-FR" xmlns="http://www.tei-
c.org/ns/1.0">Demain, dès l'aube, à l'heure où blanchit la
campagne…</quote>
```

ISO 10646/Unicode, ISO 639, ISO 3166, BCP 47, XML, TEI…

Piling up standards

Text Encoding Initiative	TEI
XML and @xml:lang	W3C
BCP 47	IETF
IANA Language Subtag Registry	IETF
ISO 639 (and ISO 15924, ISO 3166-1, UN M.49…)	ISO

ISO 639:2023, *Code for individual languages and language groups.* e.g.: fr, pt

ISO 3166, *Codes for the representation of names of countries and their subdivisions* (multipart)
e.g.: FR, PT, CA — cf. fr-FR

ISO 15924, *Codes for the representation of names of scripts*
e.g.: Cyrl, Latn — Cf. sr-Cyrl or sr-Latn

Public policies and standards - EU

European commission:

February 2022 : "An EU Strategy on Standardisation - Setting global standards in support of a resilient, green and digital EU single market"

▶ Annual Union Work Programme on Standardisation

▶ Chief Standardisation Officer

▶ EU Excellence Hub on Standards

▶ 'Standardisation Booster' to **support researchers** under Horizon 2020 and Horizon Europe to test the relevance of their results for standardisation

▶ Organise Standardisation University Days to **promote standardisation awareness** among academics and students

▶ Use the Commission's EU Academy platform for the dissemination of standardisation e-learning **training material**; promote the development and dissemination of standardisation academic teaching modules within the High-Level Forum to attract and train young professional in standardisation and promote re-skilling opportunities

Titre de la slide
(slide courante)

Limitez l'usage du gras
dans les textes courants

Premier niveau de liste (Titre)

Deuxième niveau de liste (Texte courant). Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit loborti. Mauris tempor adipiscing ligula bibendum. Vestibulum sapien lectus, porttitor vel euismod a, lobortis at mauris.

▶ Troisième niveau de liste (Puce 1). Vestibulum sapien lectus, porttitor vel euismod a, lobortis at mauris.

● Quatrième niveau de liste (Puce 2). Netus et malesuada fames ac turpis egestas.

■ Cinquième niveau de liste (Puce 2). Mauris tempor turpis eu libero sollicitudin.

Mise en page sur 2 colonnes (photo + texte)

1er niveau de liste (Titre)

Deuxième niveau de liste (Texte courant). Ut wisi enim ad minim veniam,. Vestibulum sapien lectus, porttitor vel euismod a, lobortis.

► Troisième niveau de liste (Puce 1). Vestibulum sapien lectus, vel euismod a, lobortis at.

● Quatrième niveau de liste (Puce 2). Netus et malesuada fames ac turpis.

■ Cinquième niveau de liste (Puce 3).

Mise en page sur 2 colonnes (texte + graphique)

1er niveau de liste (Titre)

Deuxième niveau de liste (Texte courant). Ut wisi enim ad minim veniam,. Vestibulum sapien lectus, porttitor vel euismod a, lobortis.
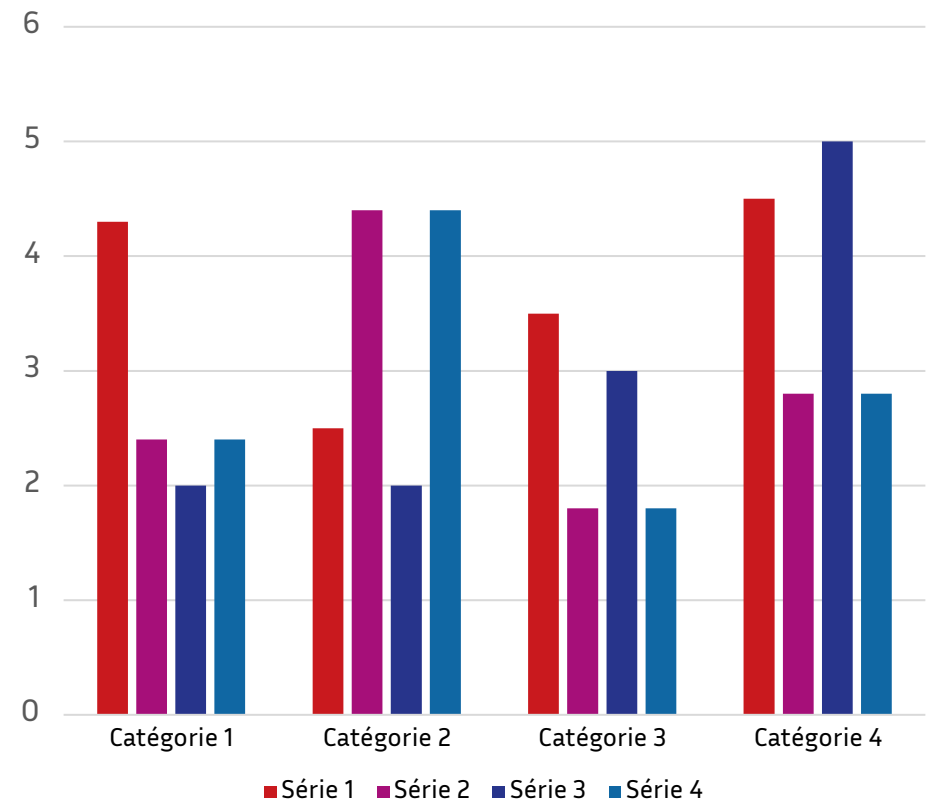
▶ Troisième niveau de liste (Puce 1). Vestibulum sapien lectus, vel euismod a, lobortis at.

● Quatrième niveau de liste (Puce 2). Netus et malesuada fames ac turpis.

■ Cinquième niveau de liste (Puce 3).

Série 1 Série 2 Série 3 Série 4

Mise en page sur 2 colonnes (tableau + texte)

	2019	2020	2021	2022	2023	2024
Janv.	0,00	0,00	0,00	0,00	0,00	0,00
Fév.	0,00	0,00	0,00	0,00	0,00	0,00
Mars	0,00	0,00	0,00	0,00	0,00	0,00
Avril	0,00	0,00	0,00	0,00	0,00	0,00
Mai	0,00	0,00	0,00	0,00	0,00	0,00
Total	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**

1er niveau de liste (Titre)

Deuxième niveau de liste (Texte courant). Ut wisi enim ad minim veniam,. Vestibulum sapien lectus, porttitor vel euismod a, lobortis.

► Troisième niveau de liste (Puce 1). Vestibulum sapien lectus, vel euismod a, lobortis at.

● Quatrième niveau de liste (Puce 2). Netus et malesuada fames ac turpis.

■ Cinquième niveau de liste (Puce 3).

Merci.

Choisir une disposition

Ajouter une diapositive (conseillé) :

Modifier une diapositive (moins performant) :