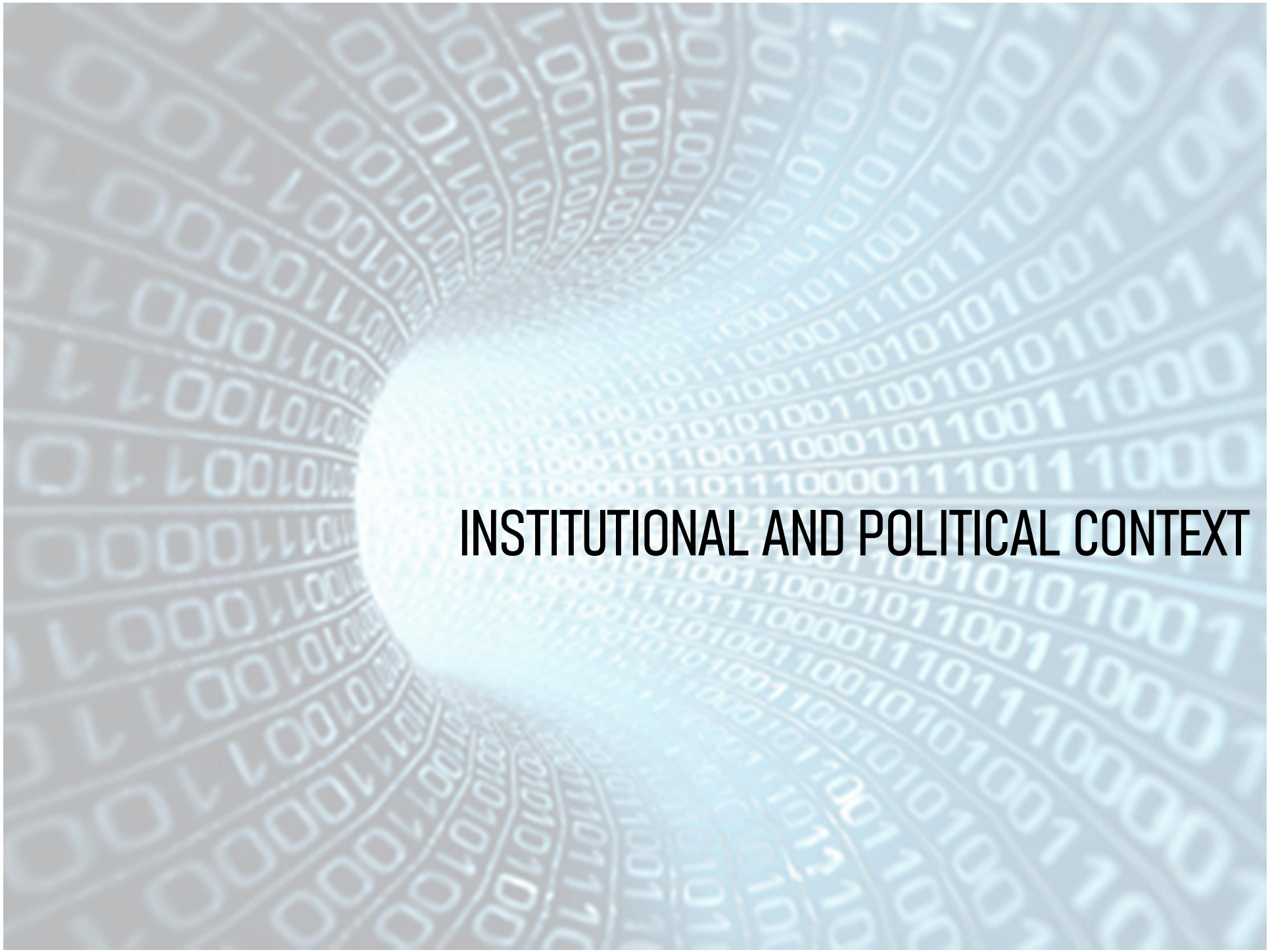
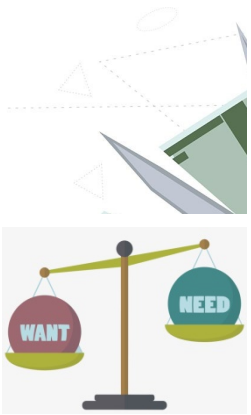

Strategies in digital archiving in the Belgian State Archives:
Between ideal and reality
Rijksarchief
Archives de l'État
Belgian Science Policy Office
belspo
Johan Van der Eycken, Belgian State Archives

Plan

- The institutional and political context
- Self-development of open source applications
- Integration of existing tools in our archival ecosystem
- Externalization to partners for development and parametrization of applications
- Externalization of some functions to federal institutional operators
- Early proactive intervention
- Conclusions
- Questions


INSTITUTIONAL AND POLITICAL CONTEXT

Institutional and political context



8 years: drastic budget cuts

New mission: collecting, preserving and disseminating digital born archives

Additionally: In 2014 we were confronted with two ageing basic tools that had to be replaced by new solutions

- State archives management system
- Search engine

no intention to extend the maintenance of these specific, tailored made tools because there were too few clients

Institutional and political context

Redefintion of competence profiles within the payroll: 2 additional webdevellopers

1,5 permanent - 1 Cont. on project
Archivists

1 Cont. on projet
Lawyer

6 agents
Experts IT

1 permanent
Engineer

1 statut. + 2 CDI + 1 Cont. on projet
Web developper

1 permanent
Webdesigner

Our strategy

Axis 1. Self-development of open source applications

Archvists make user needs and functional analysis

- IT-experts are in charge of the development
- Open source philosophy
- Applications must be compliant with the present international standards and norms
- Applications must be interoperable with the other components of our archival management
- Applications must be interoperable with external platforms and systems as the State Archives is member of different international networks like the Archive Portal Europe

Axis 2. Integration of existing tools in our archival ecosystem

- We integrate in our archival ecosystem existing tools and applications under the condition that they are open source and free**.**
- Their integration, implementation and customization to our specific needs and workflows must be realized by our own staff or by external support but in this last case, this support has to be free of charge.

Our strategy

Axis 3. Externalization to partners of the development and the parametrization of applications

- Partnership with other federal institutions or with universities or private companies that will develop applications for us or will make functional analysis
- Innovative (research) projects financed by third parties
- Disadvantage: a project means time-limited resources whose timeline doesn't fit with the obsolescence of our infrastructures and softwares
    - MADDLAIN (2015-2017) studied the expectations and needs of the digital users of our institution and of the Royal Library;
    - HECTOR (2015-2018) analyzed the management of hybrid information produced in federal public sector, and formulated advice in order to improve it and introduce best practices;
    - PROMISE (2018-2019) built a prototype for archiving the "Belgian" web. This project was coordinated by the Royal Library;
    - SODA & BISHOPS are two linked projects that aim to build an infrastructure for the preservation and the dissemination of research data produced by the Belgian research institutions
    - DIGHIMAPS: to explore the way of enhancing the geo-localization and the extraction of location names from old handwritten maps.

Our strategy

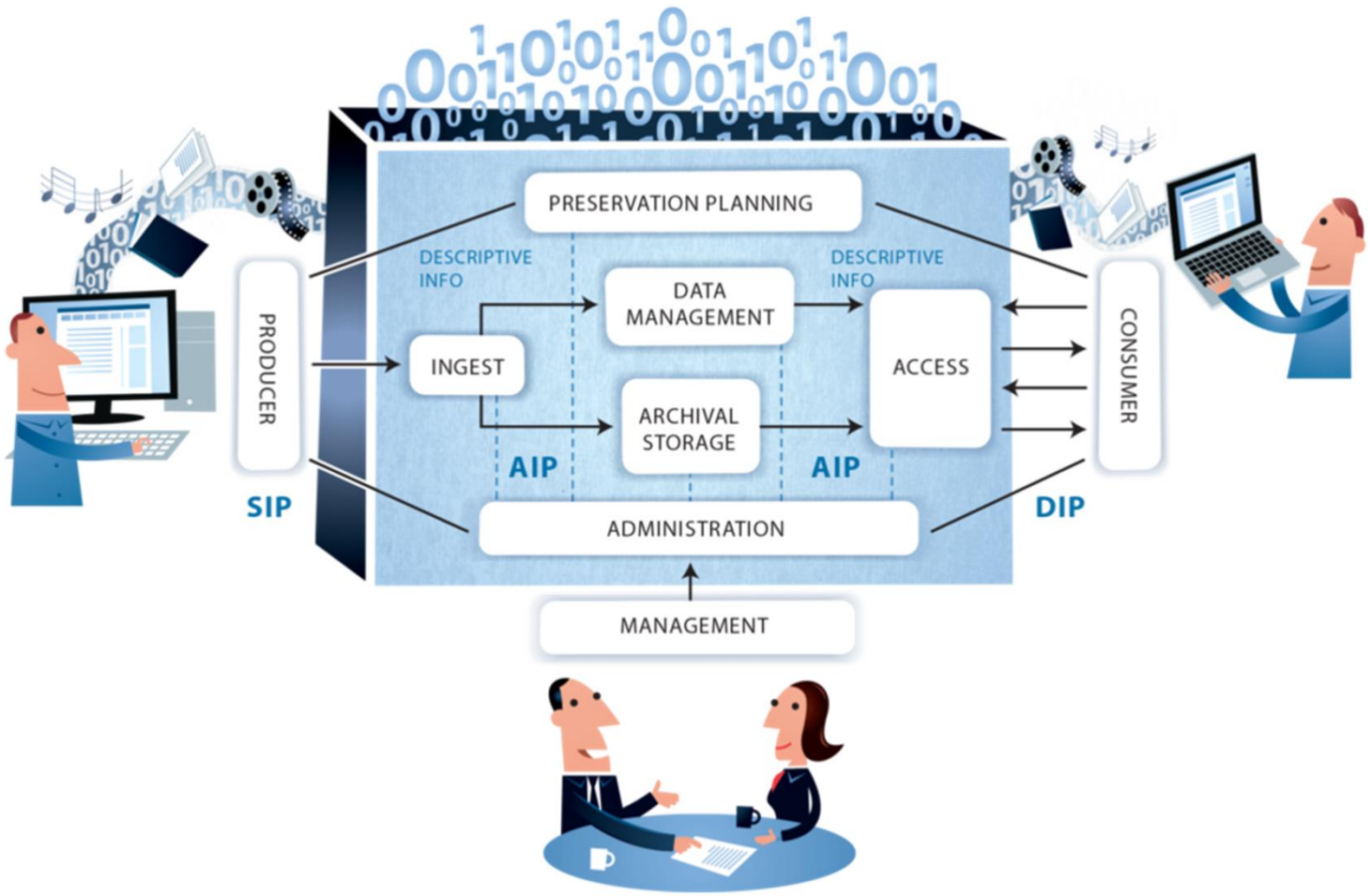
Axis 4. Externalization of some functions to federal institutional partners

The last axis consists in delegating the execution of some of our legal missions to other federal operators to whom we bring our expertise in long term archiving.

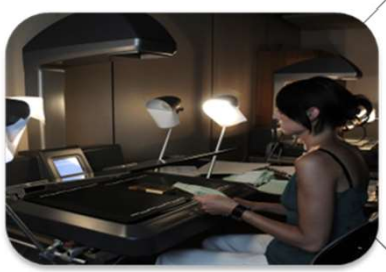
Axis 5. Promotion of records management and digital archiving in the public sector

Our last action line comprises promoting records management and digital archiving best practices in the public sector, in order to avoid the loss of digital information before the file transfer to the State Archives. Through this proactive action we hope to simplify our job when archives will be transferred to our institution.

Our archival system



Our archival system



Organize the production

- *State Archives*
- Born digital
- Scans of physical collections
- *External producers*
- Councils
- Inspection

Collecting archives

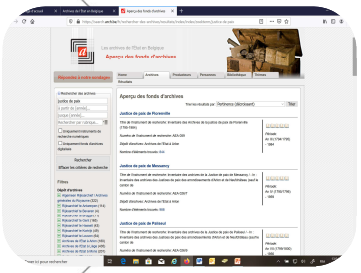
- Selection and collection policy
- Transfer procedure
- Tools
- Integrating the archives of external producers

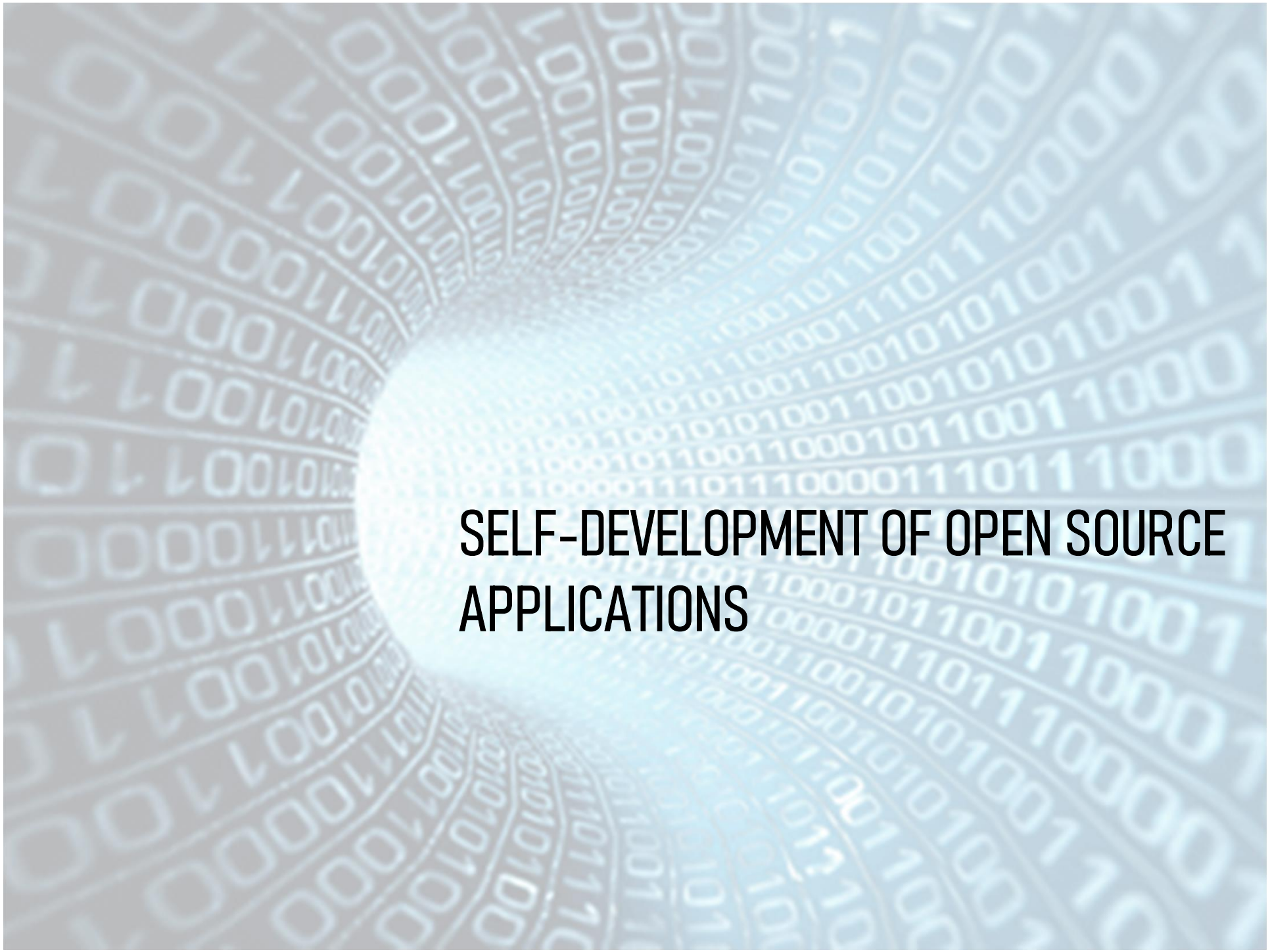
Manage and conserve

- Sustainability strategy
- For digital archives
- Managing non-digital archives with digital tools

Allow their consultation and reuse by the public

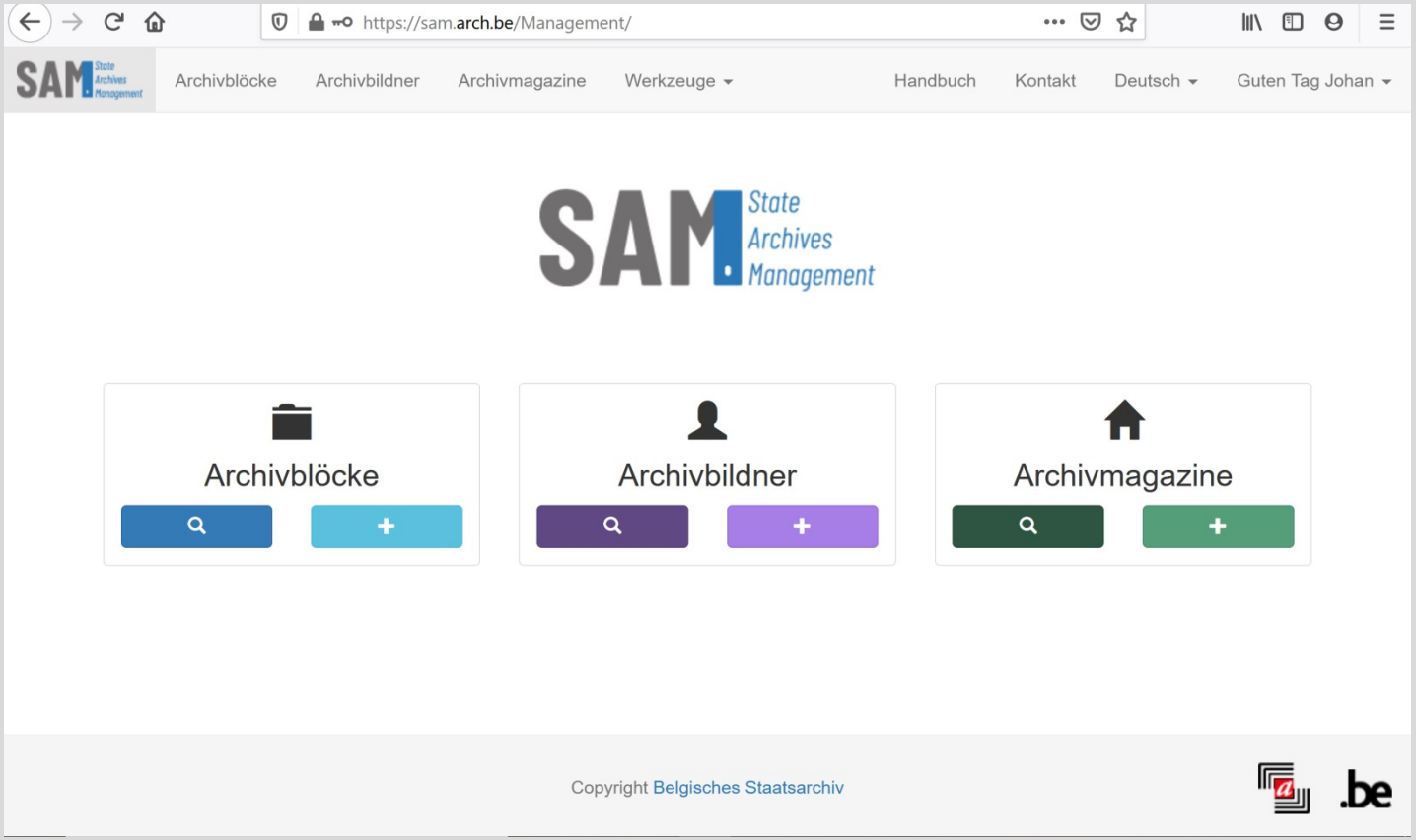
- Consultation/reutilisatio policy
- Consultation/reuse channels
- Search and visualization tools
- Delivery of copies, uploading images, etc..




SELF-DEVELOPMENT OF OPEN SOURCE
APPLICATIONS

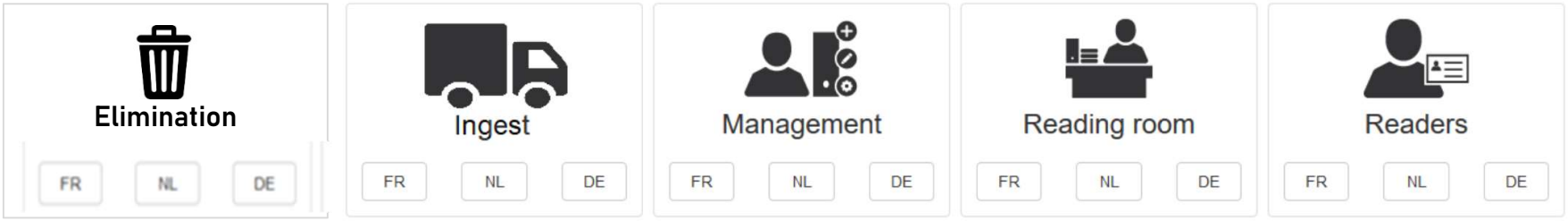
1. SAM : The State Archives Management application

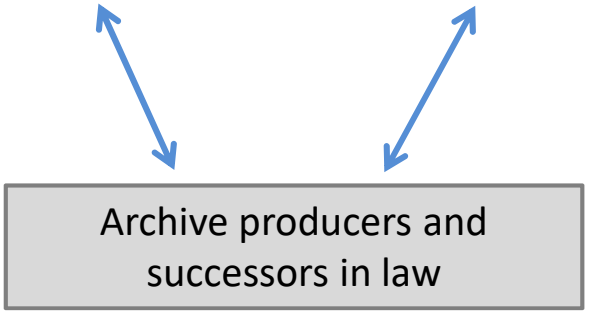
- ✓ Developed in-house, open source
- ✓ PHP, Database Maria DB, Database server + Apache server
- ✓ 3 lingual interfaces
- ✓ User friendly, simplicity and sober design
- ✓ Operational since March 2018



- ✓ Administrative metadata (storage location, format, extent, rights of access, etc.) on the digital and the analog collections are registered in SAM
- ✓ General descriptive metadata at the archive group level (title, archive producer, period, etc.)
- ✓ Series and item level descriptions are not in SAM

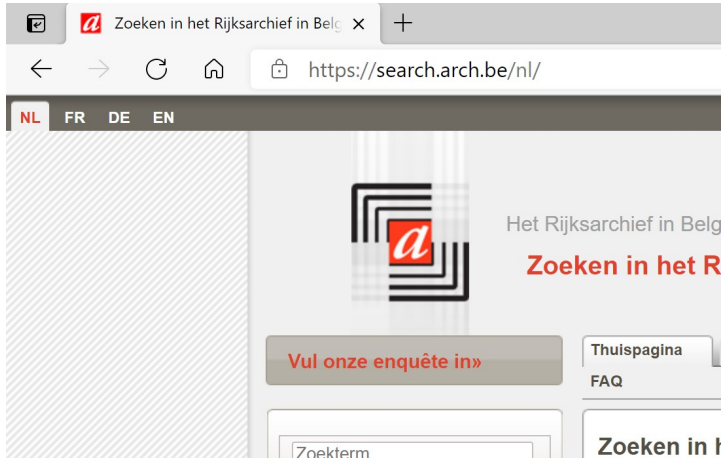
- Créer des nouveaux modules dans SAM





- Develop a workflow processing module

2. The future new search engine and website (2020-2021)



- User expectations analysis was made by the MADDLAIN project staff through an on line survey which collected about 1,000 answers.
- Face to face interviews with researchers and a workshop that brought together representatives of Belgian university researchers.
- A student trainee in Library, Archives and Documentation Sciences has completed templates for the new website and search engine design.

The replacement of the search engine will give the State Archives the opportunity to implement an Open data policy and the FAIR principles as far as possible.


INTEGRATION OF EXISTING TOOLS IN OUR ARCHIVAL ECOSYSTEM

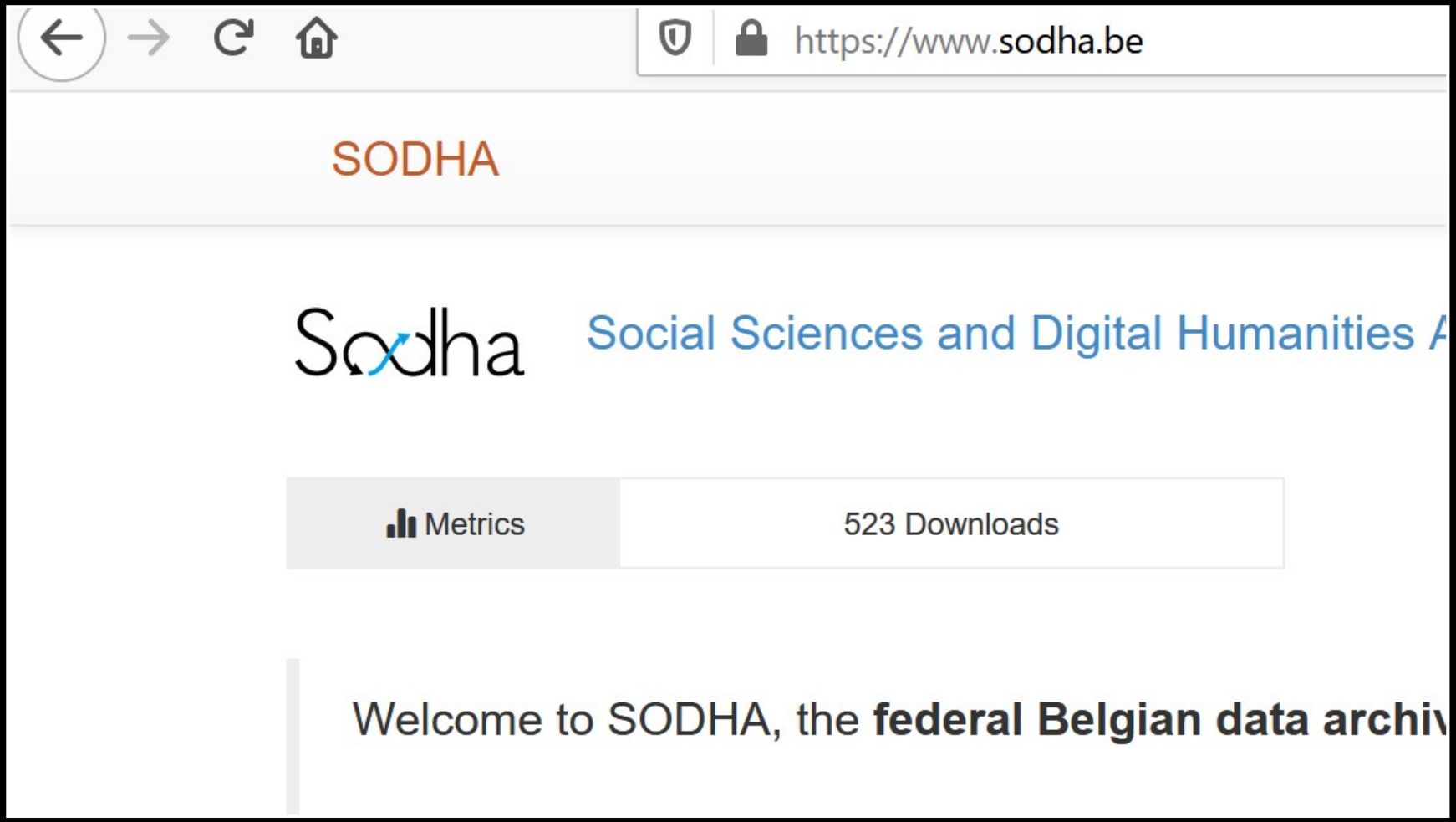
1. To transfert born digital administrative archives

- Presently our infrastructure relies on a **SFTP** and on the open source WinSCP SFTP client program that the archive producer must install on his server.
- WinSCP has been chosen because 99% of Belgian public services work with Windows. And this software is very user friendly, very simple with a drag and drop function to launch the transfer.
- We are testing some other applications especially those which allow the treatment of large bulks of unstructured office documents.
    - **Octave** (by the Archives Nationales de France)
    - **Archifiltre**
    - **Vrenamer**
    - **Bulk Rename Utility**
- → We follow closely the results of the eARK project

2. To transfer, disseminate and reuse research data

- **Dataverse** for research data:
  - Dataverse is an open source software platform ()by Harvard University)
  - It has a modular design principle using API's, that allows for distributed file storage, and supports the building of further microservices on top.
  - Dataverse has been chosen by several national CESSDA service providers like the Dutch and Austrian ones. At European level, SSHOC (the European Social Sciences & Humanities Open Cloud) is developing a data repository service for social sciences and humanities institutions, built upon the Dataverse software.
  - And Dataverse is also selected by a few Belgian universities to support their local data cross-disciplinary repository.

- Not really fit for transfer and dissemination of administrative archives.
  - The installation and customization of Dataverse requires strong IT competences.
  - No way an archival management software nor a preservation component.
    - Create the bridge between Dataverse, SAM, our own general search engine and the long term preservation infrastructure.
    - For instance we had to make a mapping between the EAD model and the DDI (Data Documentation Initiative) metadata model


https://www.sodha.be
SODHA
Sodha
Social Sciences and Digital Humanities A
Metrics
523 Downloads
Welcome to SODHA, the federal Belgian data archiv

www.sodha.be

EXTERNALIZATION TO PARTNERS FOR DEVELOPMENT AND PARAMETRIZATION OF APPLICATIONS

PROMISE: Belgian web archiving

PROMISE: Belgian web archiving

For web archiving, the PROMISE project team selected a series of softwares and combined this with some in-house development. Here a short review of the selected tools:

- To make the selection of websites we needed a tool that simplifies the creation of seed lists of websites, that must be linked to Heritrix configuration. This component is developed with Python, Django and ProstgresSQL and relies on the OCLC metadata model.
- To capture websites Heritrix has been selected as crawler, due to its ability to support broad crawls, and to the fact it is easily configurable, fast, tried and tested by several well-known cultural heritage institutions in the world. Heritrix has been compared with other crawlers Browsertrix and Brozzler. Browsertrix and Brozzler ensure high quality crawls but are slower and more experimental.
- The component providing access to the captured web archives is based on WARCLight. It is composed of a catalog that is text searchable and also searchable through a discovery function. Search criteria are the year, the public suffix, the content types, etc. The replay search function is based on PyWB and allows the search via the URL or the timestamp.
- The quality analysis tools have to be able to ensure a semi-automatic control of the visual correspondence (does the captured website look the same as the source?), the interactional correspondence (can you interact the same way?) and the completeness (do we have every resource of the original?). The tool is based on structural similarity (SSIM), visual quality indicator (VQI) and the comparison between the successful requests in archived website and the requests in original website. It balances the resource importance and takes account of factors like the content type, the CSS coverage and the image and size position.
- Derivative tools were needed in order to represent the archives in a way suited to answer some questions and analyze the archives. The Archives Unleashed Toolkit has been selected and allows to produce graphs that represent the links between websites (GraphML). With this tool it is also possible to explore the content of web documents as plain text.

EXTERNALIZATION OF SOME FUNCTIONS TO FEDERAL INSTITUTIONAL OPERATORS

Electronic archiving qualified as a G-Cloud service

As we cannot finance all the components of our digital archival system, we advocate for the adoption of shared solutions at federal level. The Belgian federal government fosters shared solutions amongst federal organizations. The range of all the solutions is named G-Cloud. The principle is that one institution develops a service and shares it with other institutions through the G-Cloud platform.

Management and conservation: storage

Aaas-platform

- shared federal records management platform
- preservation of digital semi-active records
- Internal Servers - at Smals

Financed by the State Archives

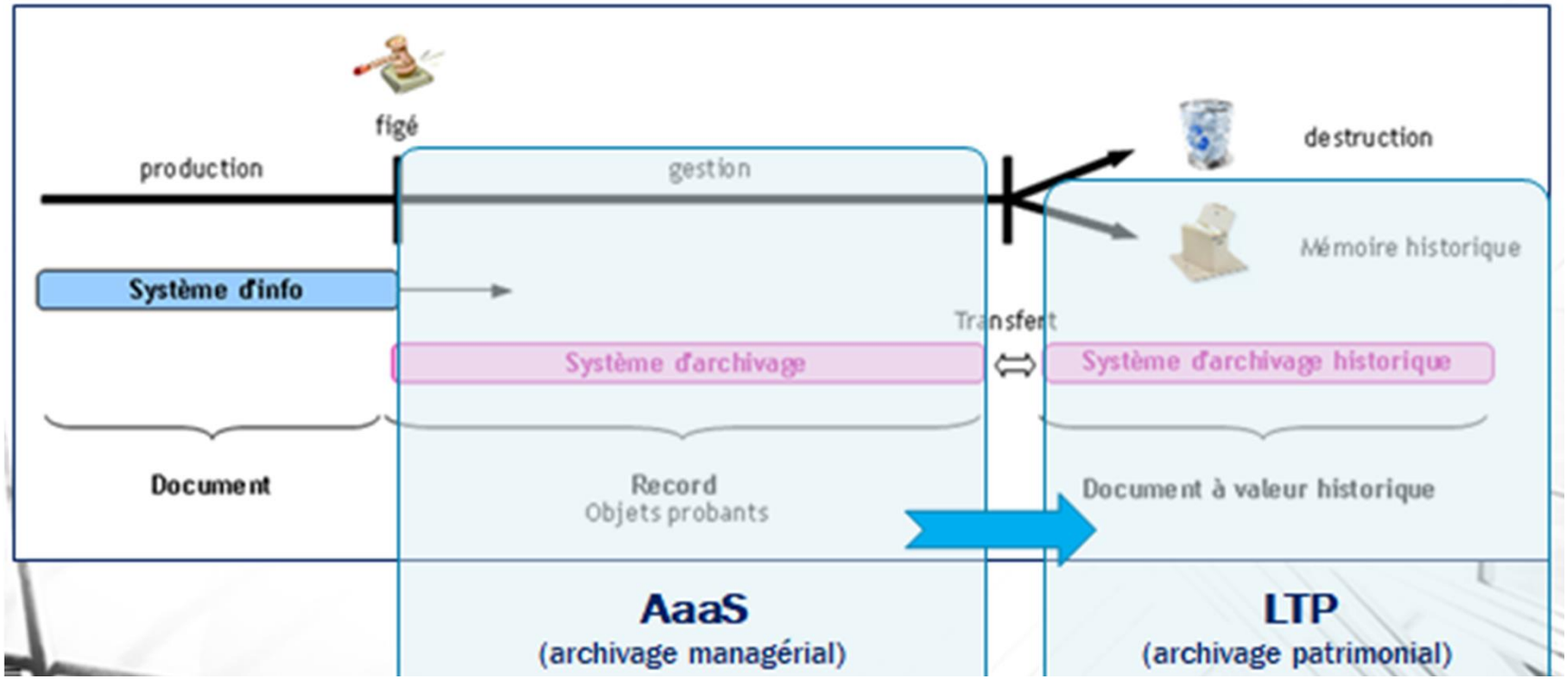
Long Term Storage Platform:

- 10 scientific and cultural heritage institutions
- compromise between all stakeholders expectations
- LTO6 tapes library

Financed by Belgian Science Policy

Promoting federal shared solutions

- Aaas et LTP

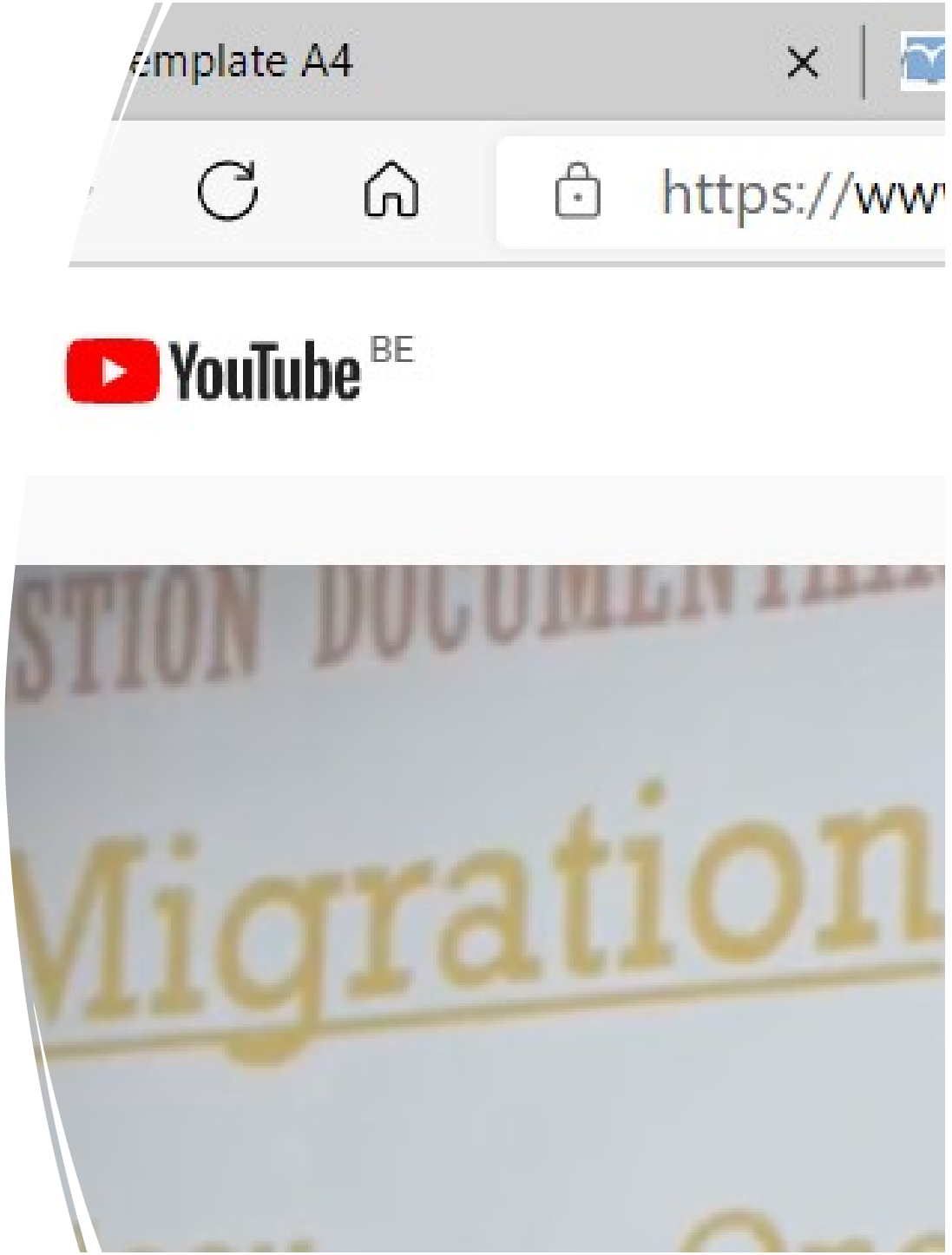



EARLY PROACTIVE INTERVENTION

Early proactive intervention

- **Advice in records keeping to civil servants**

DIMA (Digital Information Management Academy): on-the-job training methodology with workshops and offers a one year coaching during the course of a records management project (Vers un travail digital durable_Duurzaam digitaal werken – YouTube).



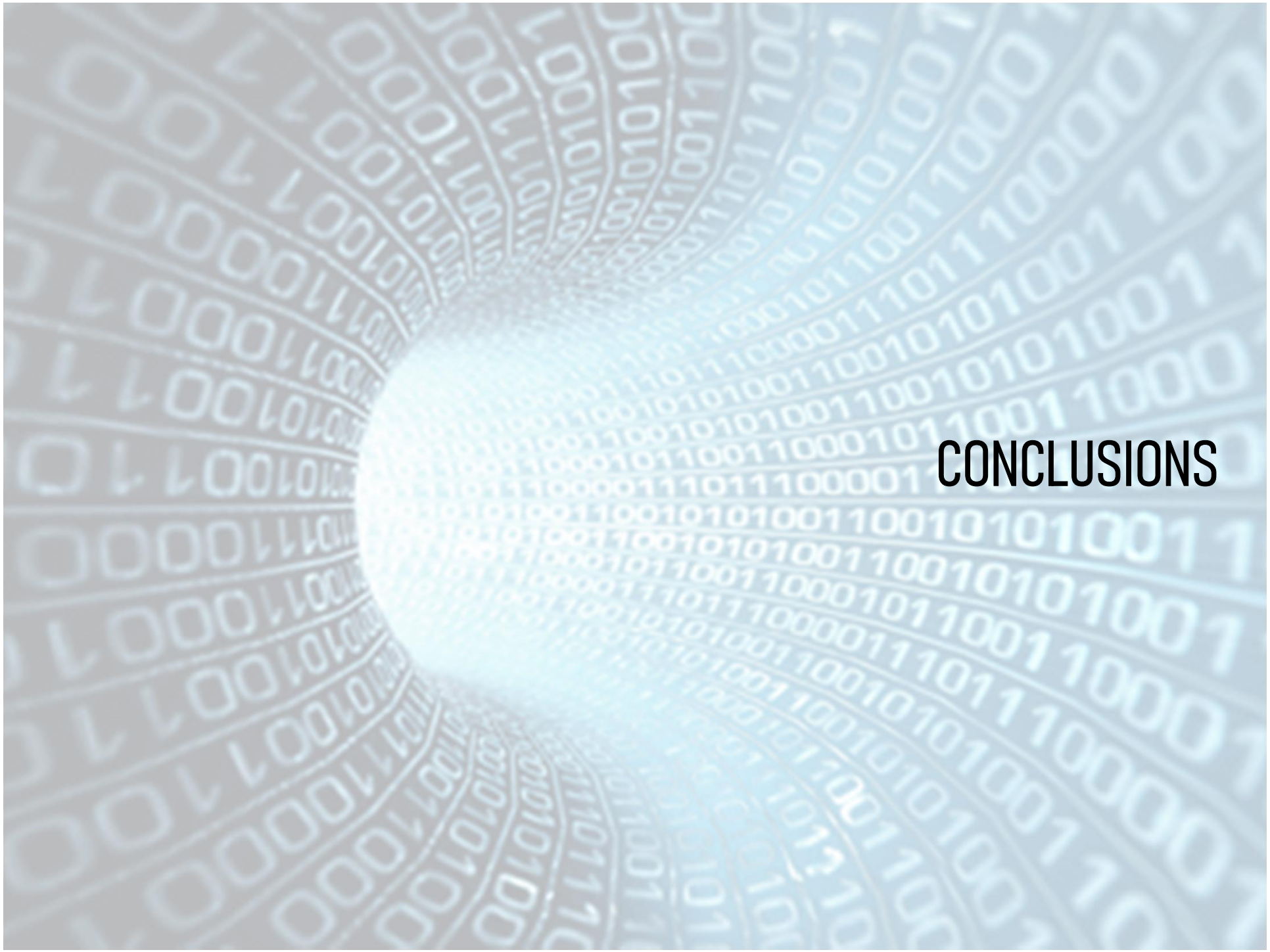
Digital Act (Law of July 21, 2016)

- **Law that regulates the use and legal consequences of electronic trust services.**
- **This law is the Belgian translation of the European eIDAS regulations.** concerns: the electronic signature for citizens; the electronic seal for companies and other legal entities; the electronic time stamp; the electronic registered mail; **electronic archiving**

Substitution of paper documents by digitization. The law creates a new framework whereby digitized documents are given the same legal value as the paper originals.

Archive producers who are subject to the Archive legislation have an obligation to keep part of their archive and to transfer it to the National Archives after the expireing of the retention period.

- In accordance with the principles of the new law, public authorities will have to use a qualified electronic archiving system for the preservation of these archives.
- The law also contains a mix of various standards and standards that must be followed.



CONCLUSIONS

Conclusions

Strengths

- The tools we get are tailor made and fit perfectly with the users expectations.
- We can build a strong and valuable internal know how that reduces our dependence to external factors and operators, for instance to private companies' commercial policies.
- The reactivity of our Development team is high and fast in case of bugs or breakdown.

Weaknesses

- The development of a solution is slow due to the small size of our technical team.
- We are dependent to the priorities and planning of partners.

Opportunities

- The delay offers the opportunity to learn from their good and less good experiences, and even to reuse their own tools.
- To potential partners we can propose our collections and our needs as case studies for research projects.

Threats

- The recruitment of IT experts is extremely difficult and slow on the Belgian job market.
- There is a risk to lose a great part of our know how when a colleague leaves the institution.

There is a need to strengthen the means

Get additional recurrent resources

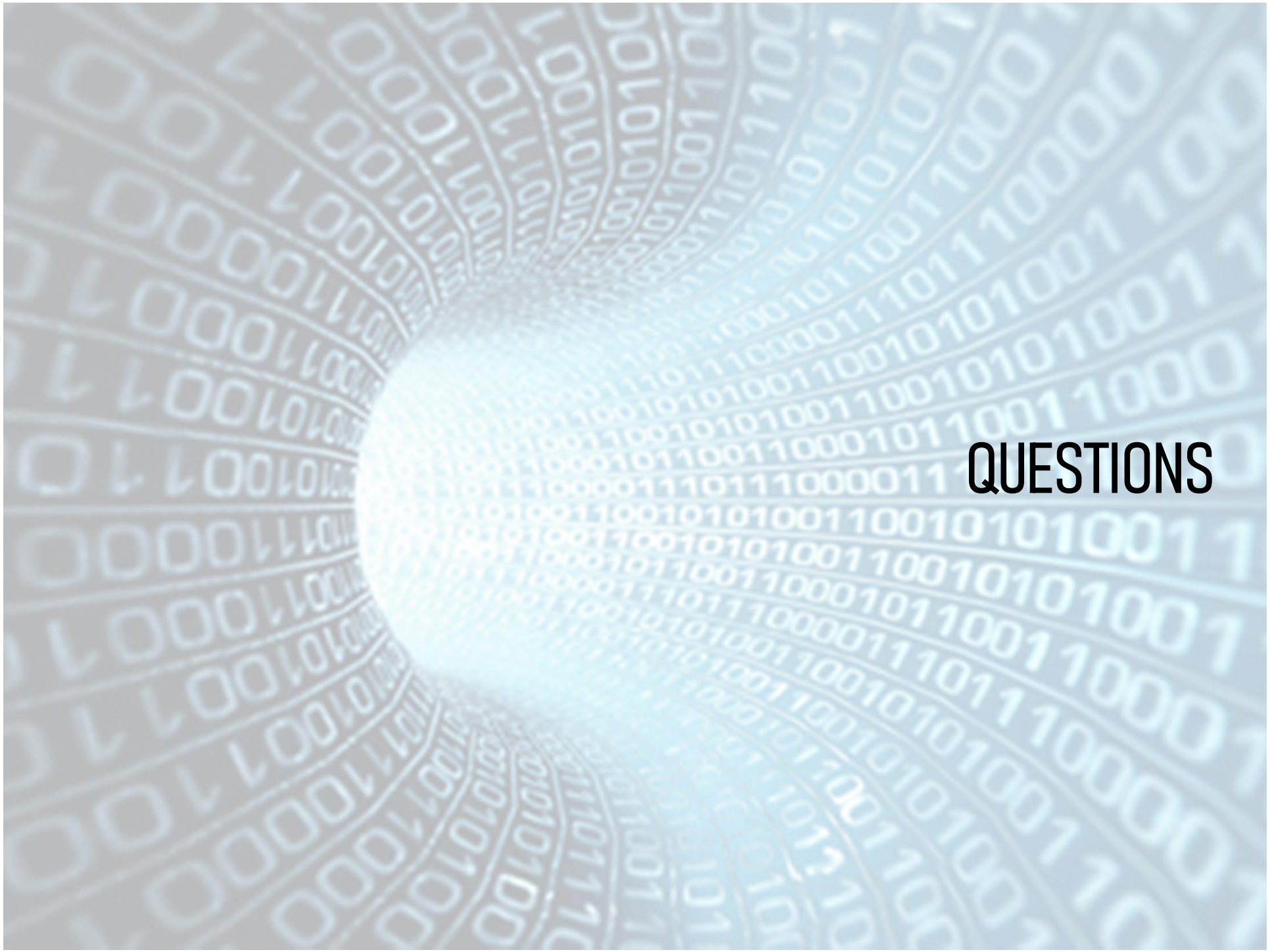
+

Promoting synergies

New profiles needed

+

Strengthening the workforce


QUESTIONS