



The Access System for Archived Databases at the DNA

Jan Dalsten, Head of Data Science

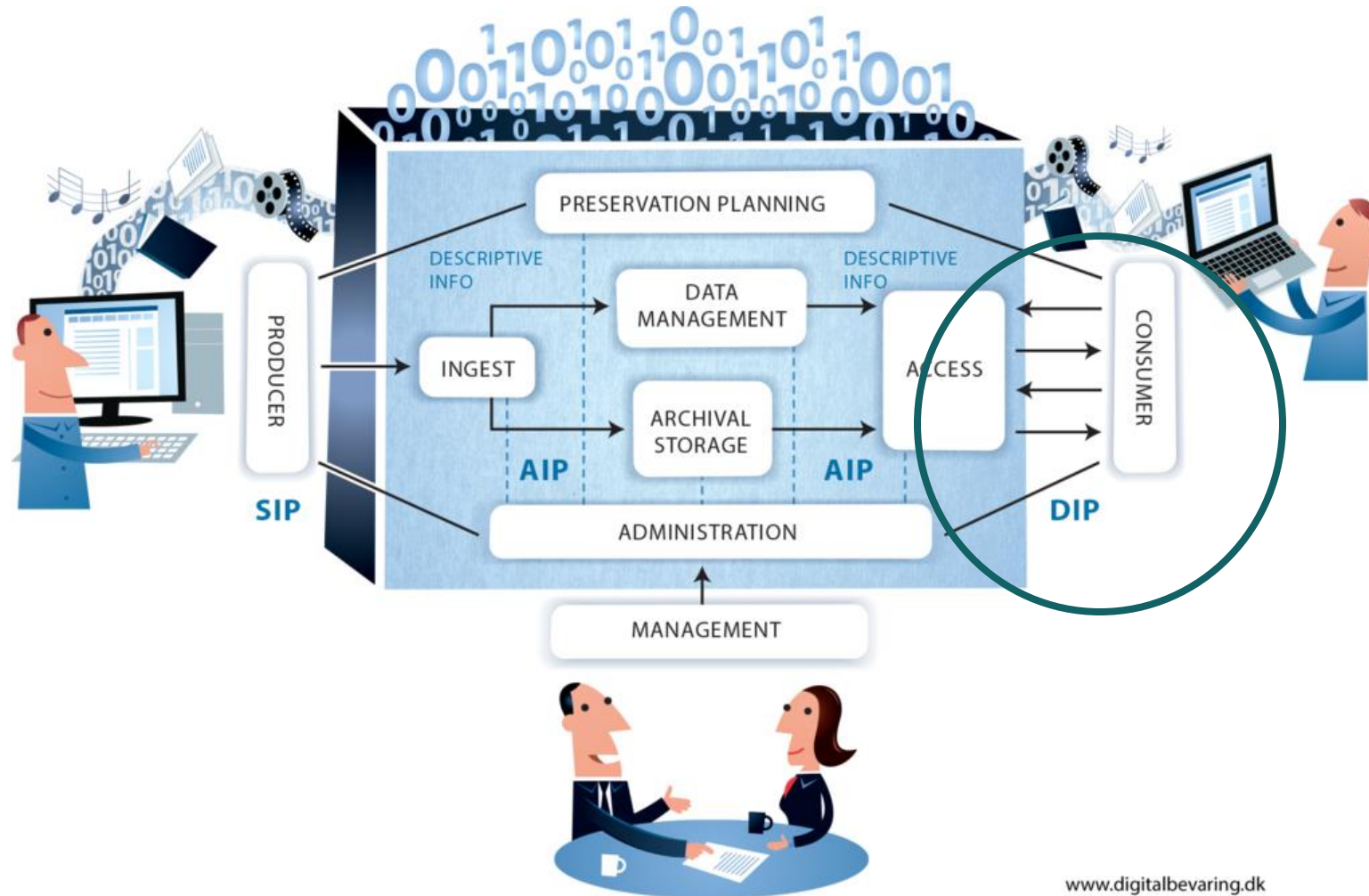
Danish National Archives

Brussels 14th May 2025

Data Science @ DNA

- Responsible for the business side of the development and maintenance of systems that give access to archived born-digital data and metadata about them
- Creation of new data sets based on paper records, either through crowd sourcing or the use of technology (Transkribus, Machine learning, AI...)
- Development of methods to link historical person data and the creation of a historical person register (HisPeR).

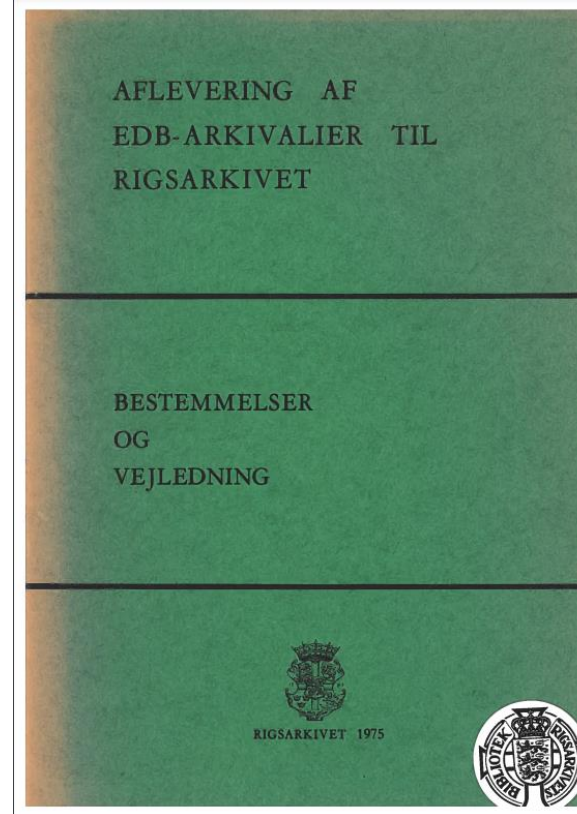
The OAIS-model



A Bit of History

- Structured regulations for SIP's since 1975
- The first access system for born digital records, "Sofia", was developed around 2008
- The amount of born-digital records is now more than 1 Petabyte (approximately 10,000 data sets)

*"Submission of
EDP-records to the
National Archives"*



*"Regulations and
guide"*

Dissemination of born-digital data at the DNA

- Dissemination is a broad term used for all the ways in which we make our data available to our users
- Two main ways:
 - Publication of metadata and freely available data (as download packages) on digidata.rigsarkivet.dk
 - Conversion from AIP's to DIP's (relational databases) for use in our search and access system Sofia

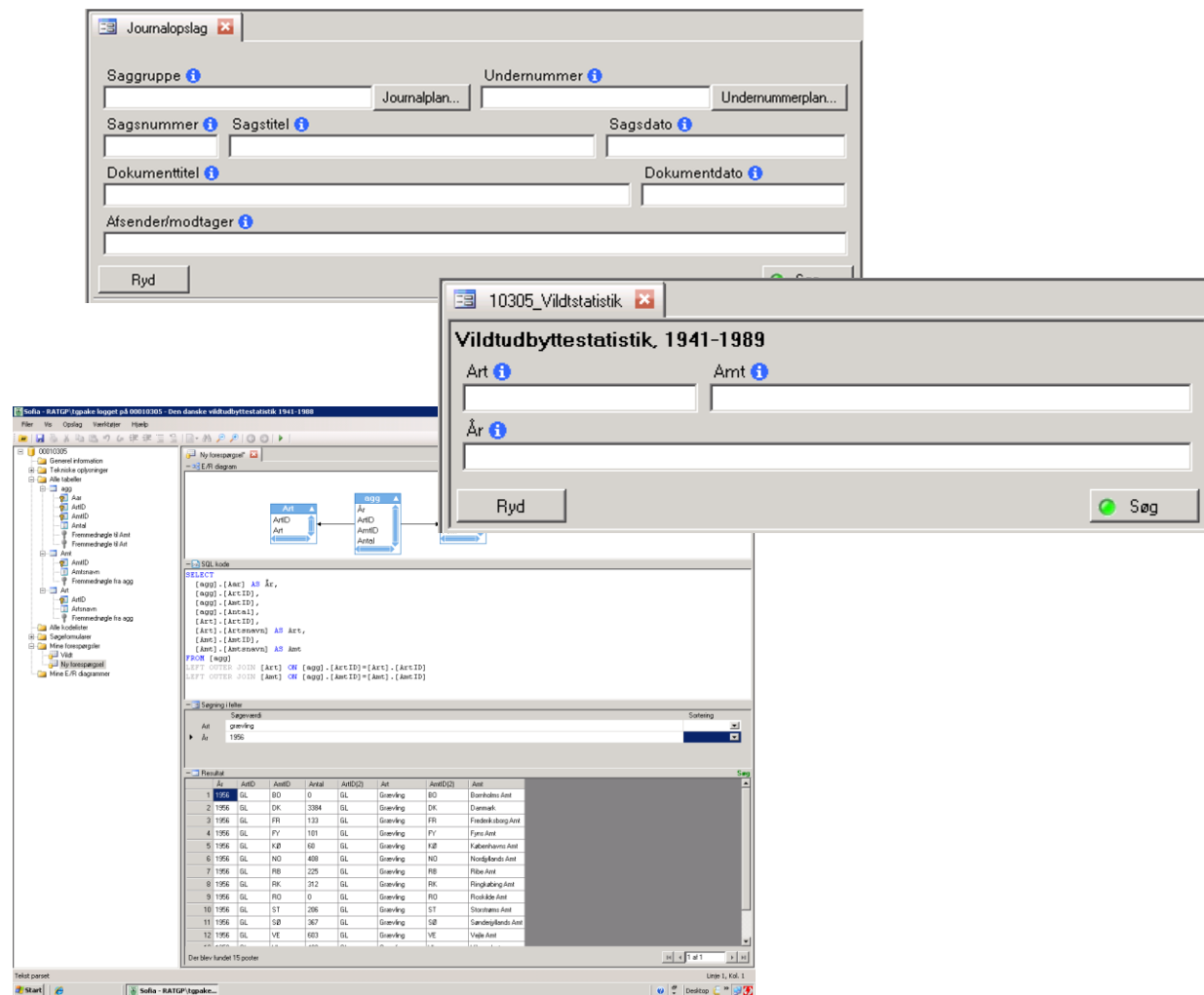
The screenshot displays the Rigsarkivet search interface. At the top, there is a search bar with the text "Hvad leder du efter?" and a "Søg" button. Below the search bar, the "Filtrér søgeresultater" section includes filters for "Datatyper" (Administrative data: 6097, Forskningsdata: 3152), "Årstal" (Fra: 1776, Til: 2023), and "Skabt af". The "Antal søgeresultater" is 9249. A message states: "Din søgning har givet 9249 resultater. Vi viser her de 100 mest relevante. Du kan præcisere din søgning ved at bruge filtrene i venstre margin." The results list includes:

- Aflevering 10006: **Befolkningsstatistikregisteret (2002)**, Skabt af: Danmarks Statistik. [Register]
- Aflevering 10: **Lærere Ved Statsseminarierne Og Ved Private Seminarier (1977)**, Skabt af: Undervisningsministeriet. [Register]
- Aflevering 10001: **Vejsektorens Informationssystem - Vejdata (1972-1996)**, Skabt af: Vejdirektoratet, Københavns kommune, Frederiksberg kommu... [Register]

Below the search results, there is a preview of a document titled "Journalopslag" with fields for "Saggruppenummer", "Sagsnummer", "Sagstitel", and "Dokumenttitel". A "Ryd" button is present. At the bottom, there is a table with the following data:

Vis:	Saggruppenr.	Saggruppetekst	Sagsnummer	Sagstitel
<input type="radio"/> Journalplanoplysninger	1	Arkiv	1	Arkiv
<input type="radio"/> Sagsoplysninger	2	dagbøger	2	dagbøger
<input checked="" type="radio"/> Dokumentoplysninger	3	erindringer	3	erindringer
	4	erindringsskitser	4	erindringsskitser
	5	ture i danmark	5	ture i danmark
	6	udenlandsrejser	6	udenlandsrejser

Searching in Sofia

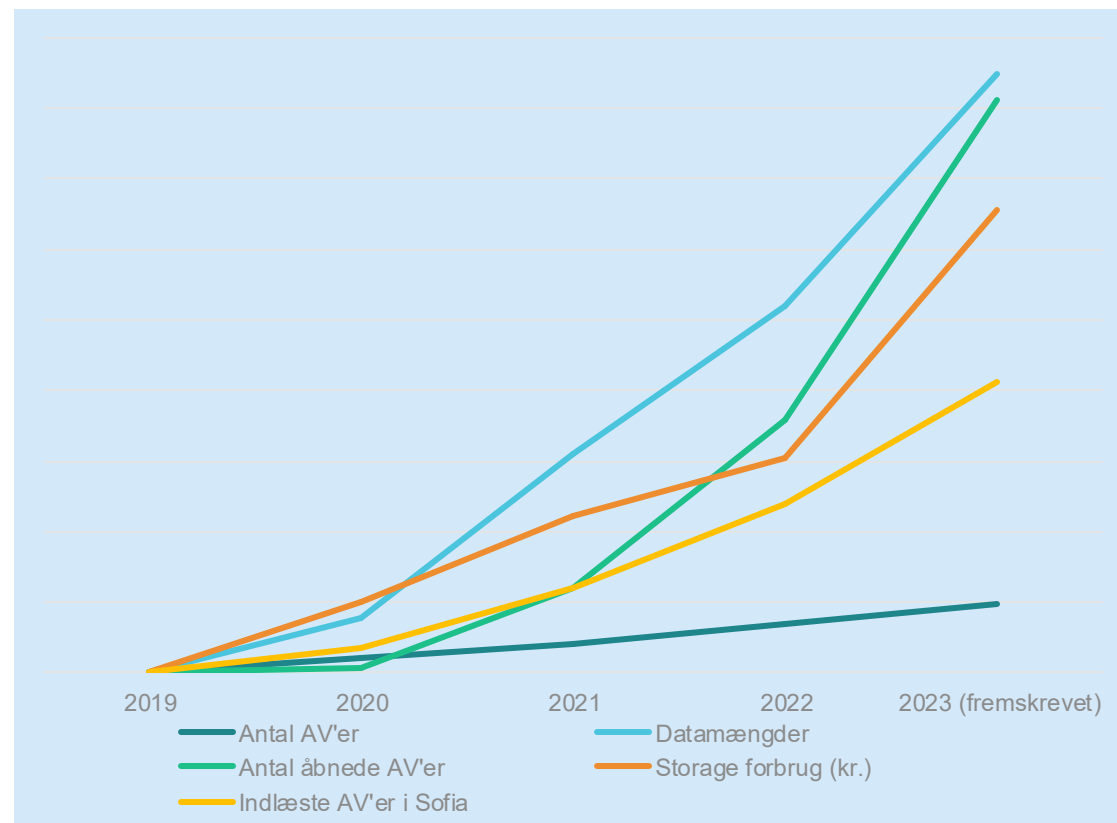


Sofia has three different search templates:

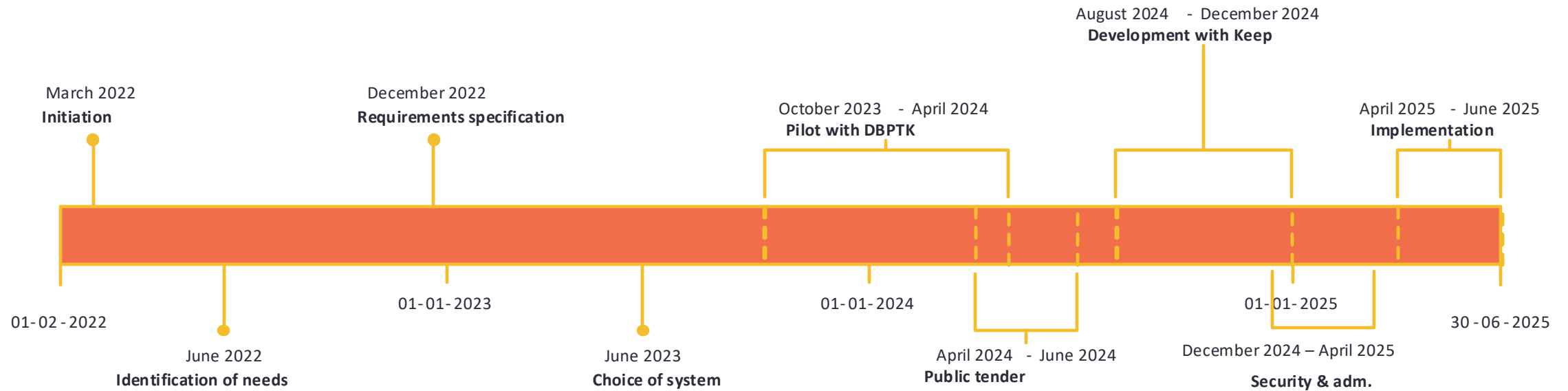
1. Records management systems for AIPs with documents
2. Other kinds of digital records for AIP's without documents
3. SQL-queries

Why did we need a new access system?

- We have implemented SIARD (in a Danish variety) as preservation format for databases and Sofia has only been partially adapted to this
- Sofia had reached end of life and had become really costly to maintain
- The use of Sofia requires manual pre-processing of data (e.g. mapping to file name, document name etc.)
- The requests for information have gone up considerably, and Sofia and the infrastructure around it could not scale sufficiently



The process



Choice of system

- Develop a system ourselves or buy "off the shelf"?
- It is costly to develop yourself, but there was no system on the market that could fulfill all our requirements, including:
 - Faster (and more automatic) load and indexing of data
 - Better search facilities
 - Management of user access and user rights
 - Better support for LOBs
 - Better possibilities for the export of data and documents



Database Preservation Toolkit (DBPTK)

- DBPTK is a collection of multiple tools developed in relation to the E-ARK project
- DBPTK consist of two primary tools: a SIARD-creation tool and a SIARD-viewer, where it is possible to search through data in a "Google-like" manner
- DBPTK exists in three editions: Desktop, Enterprise and Developer



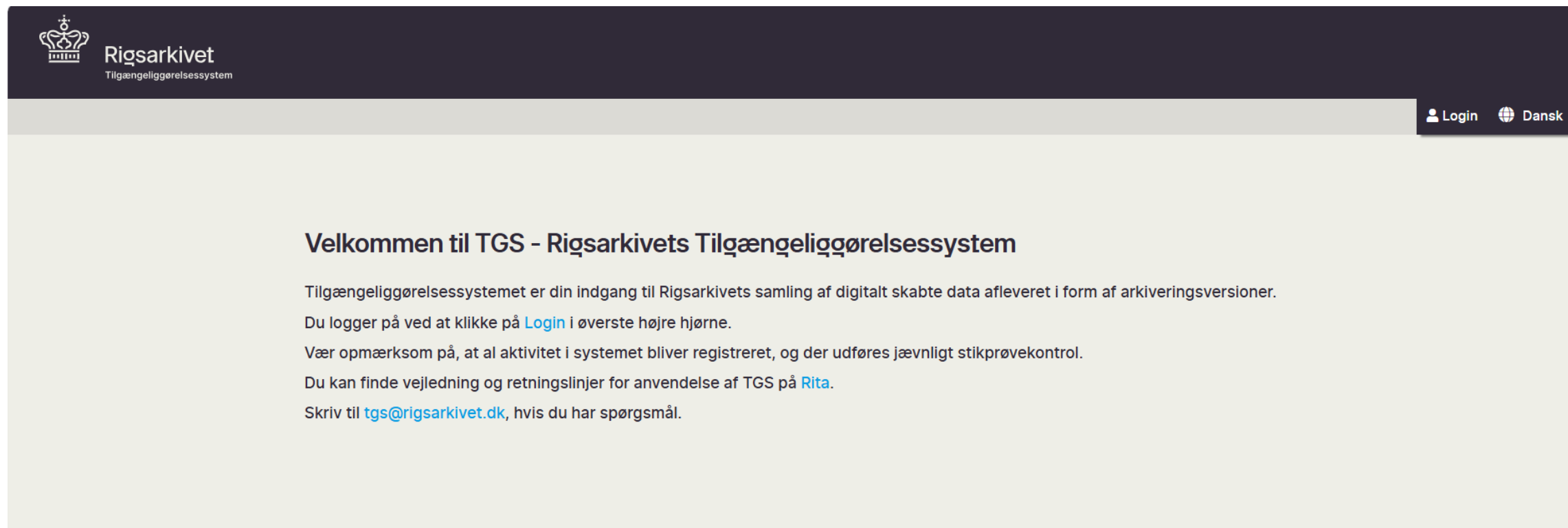
database^{toolkit}
preservation



Necessary developments

- DBPTK met \approx 80 pct. of our requirements "out-of-the box"
- Identified needs for development:
 - Support for SIARD DK
 - User management
 - Integrated document viewer (TIFF, sound, video)
 - Search across databases
 - Translation to Danish

DNA "look and feel"



Why such a long process?

- Important to make sure that the business needs were identified
- Regulations regarding procurement
- Increased focus on GDPR, data protection etc. made a revision of the security set-up necessary

Some final thoughts...

- It is difficult to find enough internal developer resources to maintain all our systems
- The European cooperation in the E-ARK-project (etc.) has paved the way for our ability to use external developers for archival core systems
- The use of standards such as OAIS and SIARD is really helpful!
- Open Source reduces the risk of vendor lock-in

Thank you



© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

