Enabling metadata driven Research data platform: transforming Static metadata to Smart metadata

Archana Bidargaddi

Morten Jackobsen

Sikt – Norwegian Agency for Shared Services in E & R

Conference on Smart Metadata for Official Statistics (COSMOS)

11-12 April 2024, Paris

Sikt

Surveybanken

Sikt

The Problem with data

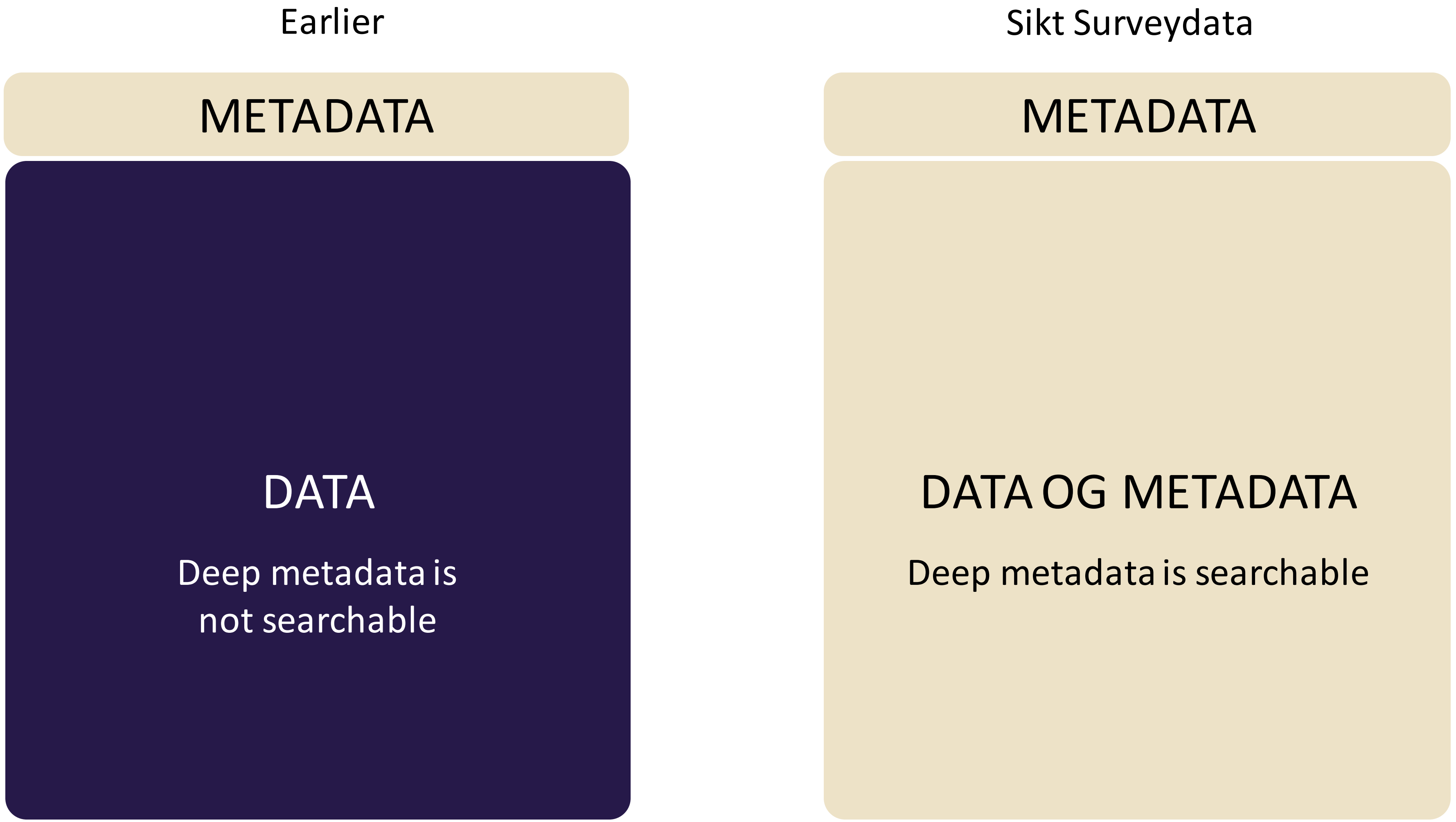
F – unable to Find

A – difficult to Access

I – not machine actionable

R – not sufficient metadata

Easier to find and reuse research data



Earlier

METADATA

DATA

Deep metadata is not searchable

Sikt Surveydata

METADATA

DATA OG METADATA

Deep metadata is searchable

Smart metadata driven

Search

Visualisation

Analysis

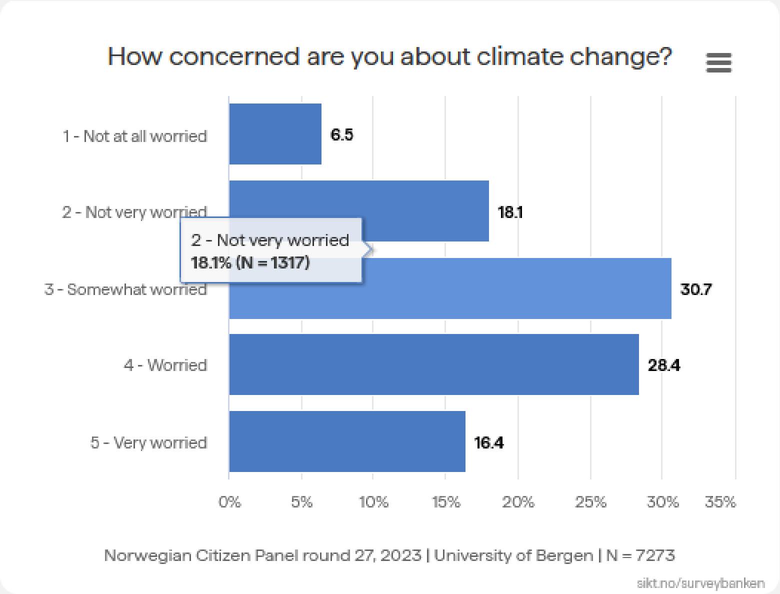
Data access

Surveybanken

How concerned are you about climate change?

Example of automatic visualization of answers to the how concerned one is about climate change. The answers are automatically distributed according to background variables defined in metadata such as gender, age and education.

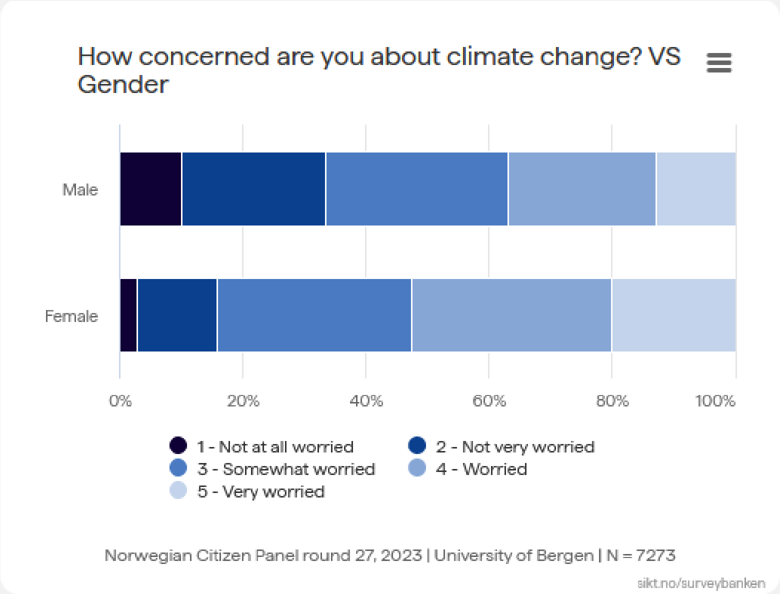
https://surveybanken.sikt.no/en/study/8776be13-fd70-4cc0-8df1-e85803ac36d2/14?type=analyses&elements=[%229eed47f1-3314-422f-b0d3-34889a708ba8/2%22]&datafile=f26b538c-553d-47a6-aec2-01bb4846290d/11



How concerned are you about climate change?

Norwegian Citizen Panel round 27, 2023 | University of Bergen | N = 7273

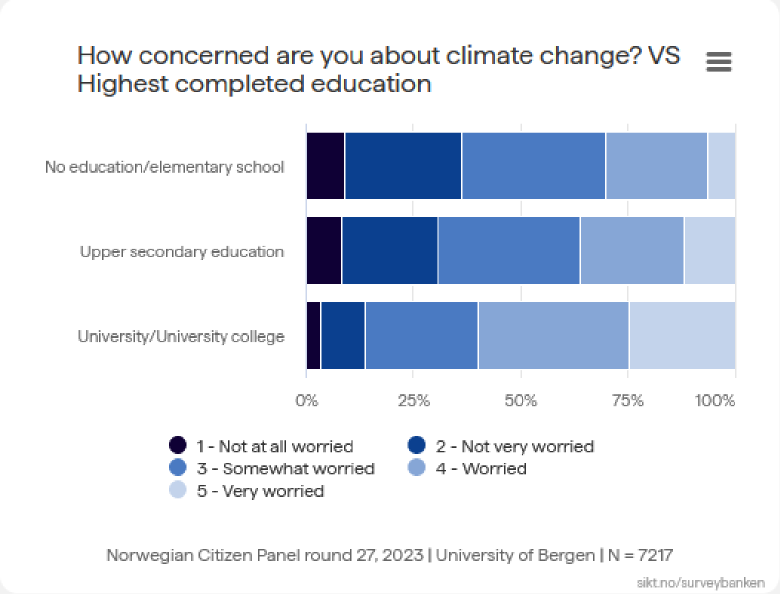
sikt.no/surveybanken



How concerned are you about climate change? VS Gender

Norwegian Citizen Panel round 27, 2023 | University of Bergen | N = 7273

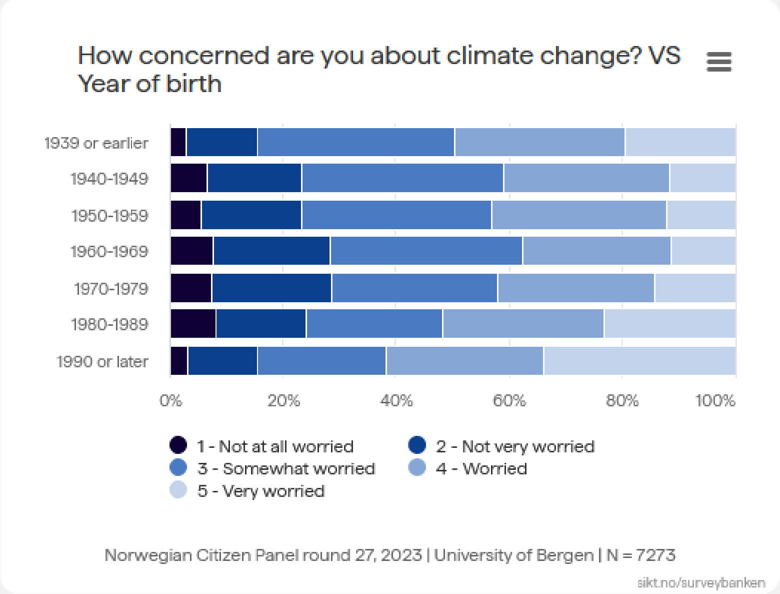
sikt.no/surveybanken



How concerned are you about climate change? VS Highest completed education

Norwegian Citizen Panel round 27, 2023 | University of Bergen | N = 7217

sikt.no/surveybanken



How concerned are you about climate change? VS Year of birth

Norwegian Citizen Panel round 27, 2023 | University of Bergen | N = 7273

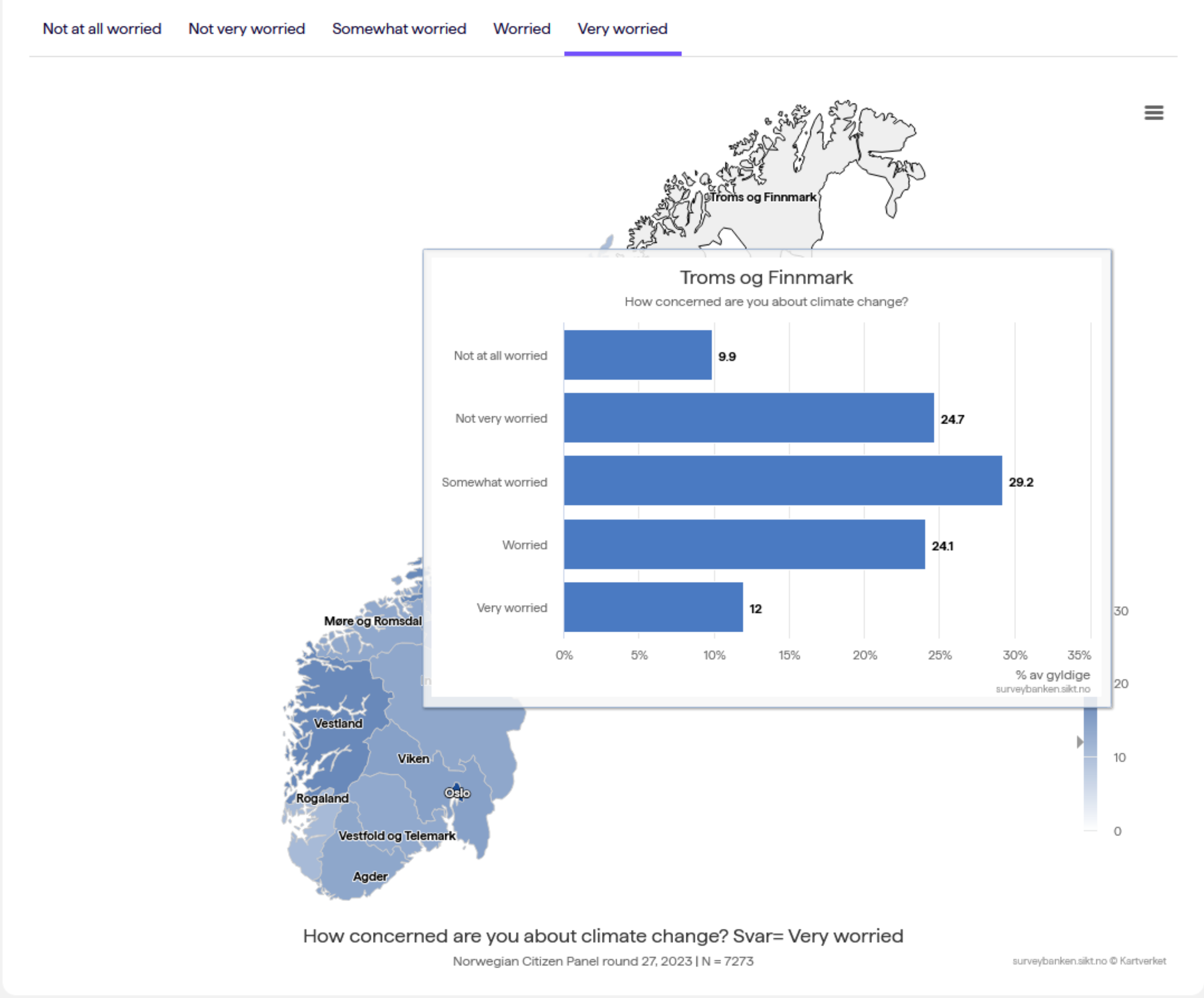
sikt.no/surveybanken

Sikt

How concerned are you about climate change?

Example of automatic map visualisation of answers to the how concerned one is about climate change. Selection of the relevant map to display is driven by metadata.

https://surveybanken.sikt.no/en/study/8776be13-fd70-4cc0-8df1-e85803ac36d2/14?type=analyses&elements=[%229eed47f1-3314-422f-b0d3-34889a708ba8/2%22]&datafile=f26b538c-553d-47a6-aec2-01bb4846290d/11



Troms og Finnmark

How concerned are you about climate change?

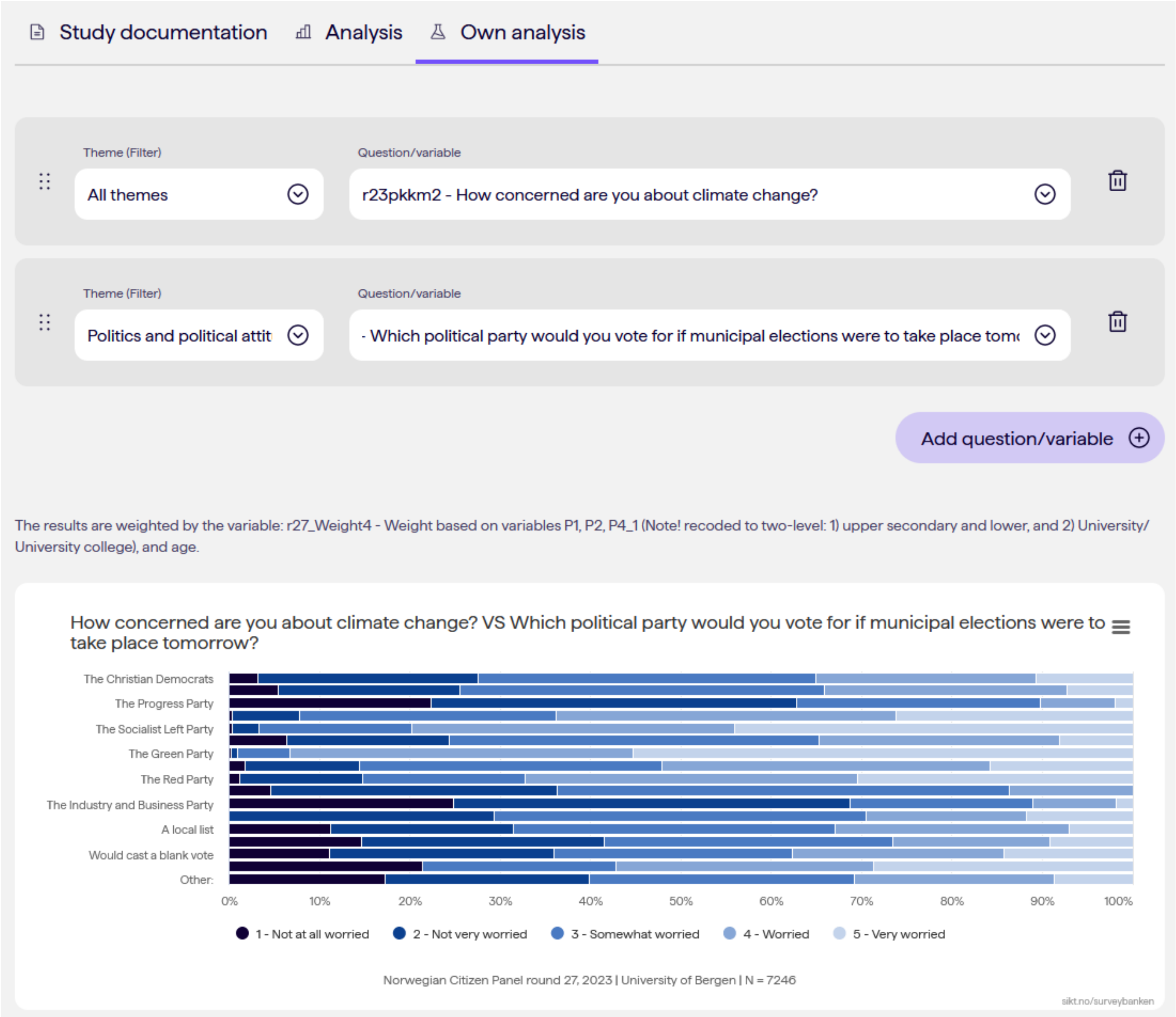
How concerned are you about climate change? Svar= Very worried

Norwegian Citizen Panel round 27, 2023 | N = 7273

How concerned are you about climate change?

Example of own analysis user can perform to explore answers to how concerned one is about climate change. Selection of the relevant visualisation to display is driven by metadata.

https://surveybanken.sikt.no/en/study/8776be13-fd70-4cc0-8df1-e85803ac36d2/14?type=ownAnalyses&elements=[%229eed47f1-3314-422f-b0d3-34889a708ba8/2%22,%22e0dcd24c-0104-4b7f-9e1b-ef4c0fbf0691/2%22]&datafile=f26b538c-553d-47a6-aec2-01bb4846290d/11



Theme (Filter): All themes

Question/variable: r23pkkm2 - How concerned are you about climate change?

Theme (Filter): Politics and political attit

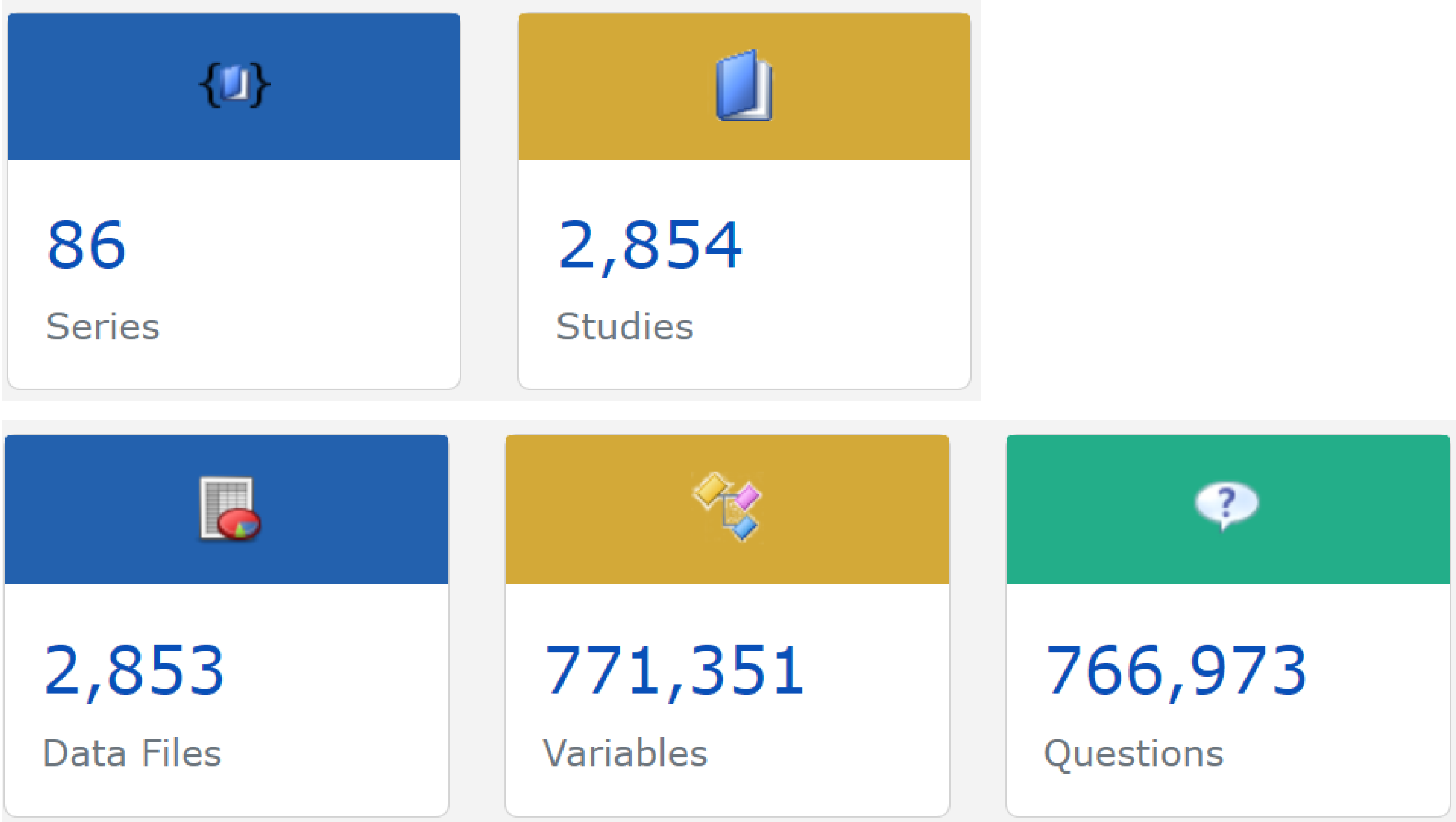
Question/variable: - Which political party would you vote for if municipal elections were to take place tom

The results are weighted by the variable: r27_Weight4 - Weight based on variables P1, P2, P4_1 (Note! recoded to two-level: 1) upper secondary and lower, and 2) University/University college), and age.

How concerned are you about climate change? VS Which political party would you vote for if municipal elections were to take place tomorrow?

Norwegian Citizen Panel round 27, 2023 | University of Bergen | N = 7246

Norwegian Research Data Archive



86

Series

2,854

Studies

2,853

Data Files

771,351

Variables

766,973

Questions

Sikt's data archive legacy infrastructure

Shared file storage on LAN - SIP, Nesstar files, related materials

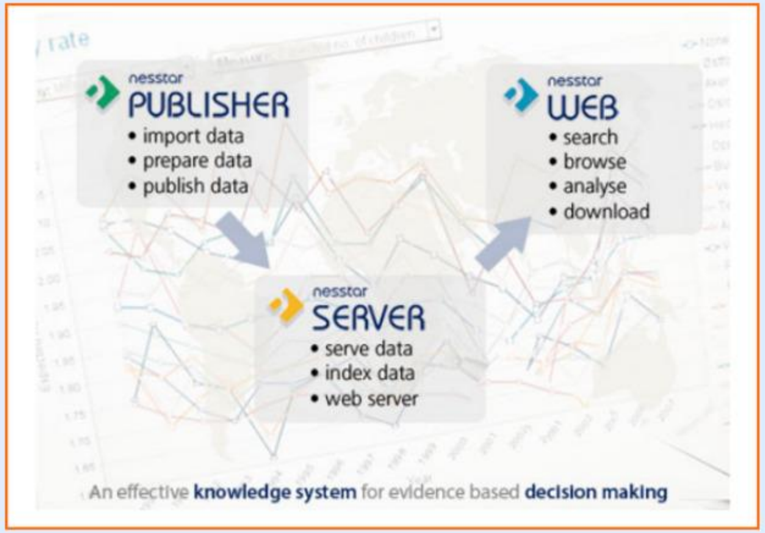
SPSS

Statiscal software for data processing

StudyNum - Master-list of studies

REDMINE

Ticketing System



nesstar PUBLISHER
- import data
- prepare data
- publish data

nesstar WEB
- search
- browse
- analyse
- download

nesstar SERVER
- serve data
- index data
- web server

An effective **knowledge system** for evidence based **decision making**

StudyAdmin

Internal curation notes

Email based communication

Nesstar: Metadata documentation and dissemination software

FileSender

Data delivery

Background

- 20+ instances of Nesstar
- Over 2500++ studies
- 90% documented at variable level
- Manual data processing in SPSS
- Communication via email
- Redmine ticketing system used for both data processing and data access
- Internal documentation held separately

Sikt

So then what?

Sikt

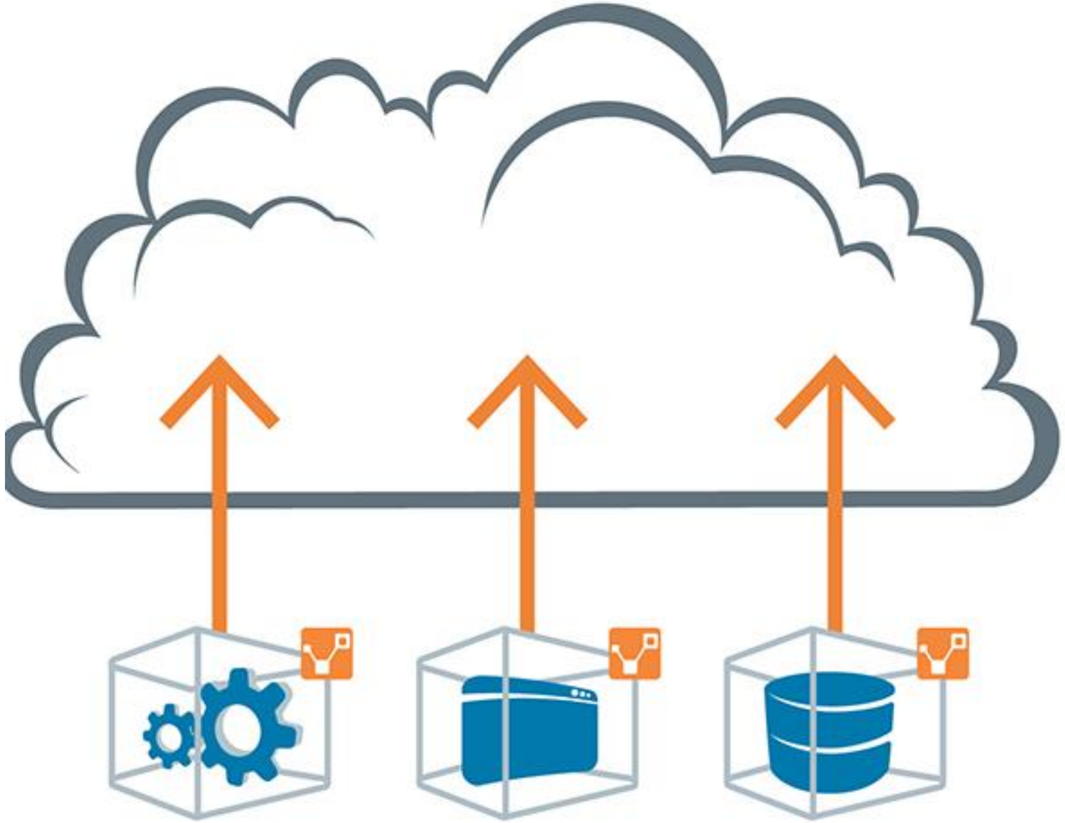
Major infrastructure modernisation projects

From On-premise infrastructure to Cloud Infrastructure

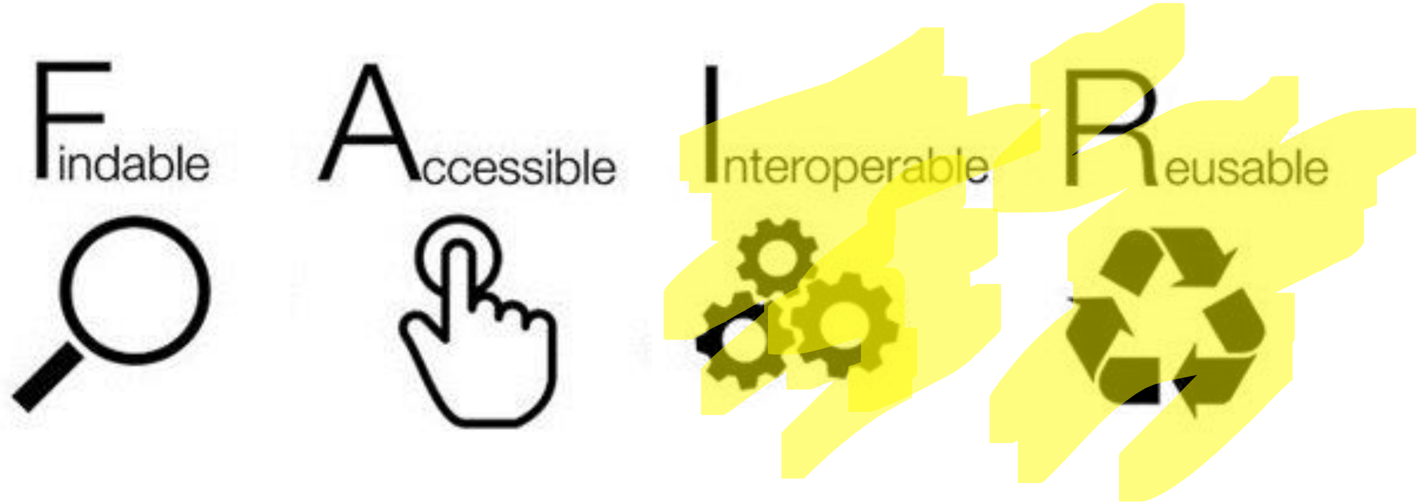
- Old technology
- Disintegrated systems
- Non-collaborative services

To:

- Modern technology
- Integrated systems
- Consolidation of solution
- Cleaning of data and metadata
- Collaborative services



Sikt


Findable
Accessible
Interoperable
Reusable

Transforming Static metadata to Smart metadata

Sikt

Goals

- The curators at the archive should have more control of the metadata they manage
- Shift the authority of the information from the single record to the centralized resource package
- Establish a single source of truth for metadata elements
- Foundation for building metadata-driven systems

Sikt

Process

- Cleaning and harmonisation of metadata
- DDI-Lifecycle metadata profile
- Transforming metadata to be machine-actionable
- The migration process
- Enabling metadata driven services
- Connecting data and metadata

Metadata quality in legacy system

- DOI in a custom metadata element
- Typos in metadata entry
- Many custom elements
- Manual versioning of the Nesstar-files
- Many instances of free text in place of Controlled Vocabularies and controlled lists
- Subjective tagging av keywords from ELSST

Cleaning and Harmonization

- Information in elements supporting controlled vocabularies were mapped to relevant descriptive terms and code values extracted from CESSDA Vocabulary Service.
- Aggregation to top levels for organisations – from 3000 to 300
- Access conditions mapped to were mapped into 6 broad categories ranging from open access to restricted personal data – from 2700 to 6
- Flatten variable groups
- Dropped all keywords as cleaning would be effort-intensive

Sikt

DDI-Lifecycle Metadata Profile

Considerations:

- CESSDA Metadata Model had 500 elements
- What was supported by Colectica
- What we had in DDI-Codebook and wanted to carry over
- And what we needed

Results:

- A subset of CESSDA Metadata Model
- Upgrades to Colectica:
  - controlled vocabularies and
  - how the DDI elements Creator, Contributor and Publisher were structured.

Sikt

Transforming metadata to be machine-actionable

- Defining Key Elements as Resource Packages
  - Concept schemes, Organization schemes, Code lists, Archives, Groups and Other material
- Mapping study elements to the information in the Resource Packages on import.

Sikt

Metadata Migration – agile and iterative process

- Harmonization mapping files for cleaning
- Automated batch cleaning of Nesstar files
- Mapping elements upon batch import into Colectica
- Review metadata after import
- Finetune "Metadata profile", "Harmonization mapping files" and batch processes based on review
- Repeat export and import

Enabling metadata driven services

Depending on which key metadata element is referenced by the single record, systems have stable information on how to display, make available, and analyze nuanced digital objects.

Sikt

Enabling metadata driven services

- Versioning of data and documents
  - MAJOR - inclusion of new country data
  - MINOR - Changes in data or metadata that will influence the use of data or the results of data analysis
  - PATCH - insignificant changes such as spelling errors
- Versioning of metadata
  - Every version of an item committed to the repository is saved, allowing clients to retrieve a full version history of any item in the repository.
- Controlled Vocabularies
  - Increased interoperability was achieved by implementation of controlled vocabularies and statistical classifications
- Universal Unique Identifiers
  - Every metadata element has an UUID, enabling systems to identify and retrieve the element.

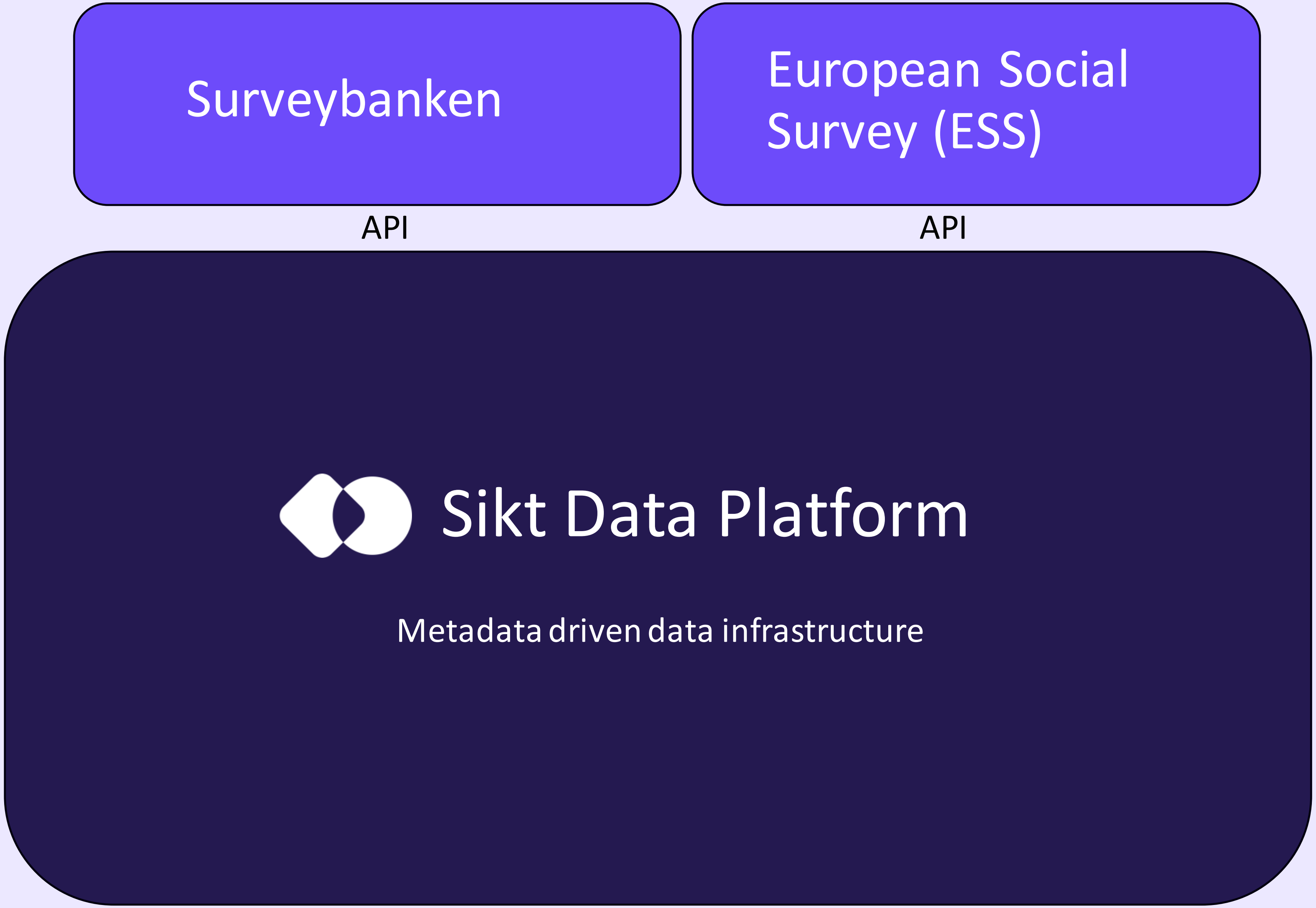
Connecting data and metadata

- Versioned metadata elements
- Immutable, flat-file for data storage
- Reference data in metadata
- PhysicalInstanceId+version drive data download and analysis
- Data processing APIs to calibrate data with metadata real-time

Sikt

The new metadata driven Research Data Platform

Sikt



Surveybanken

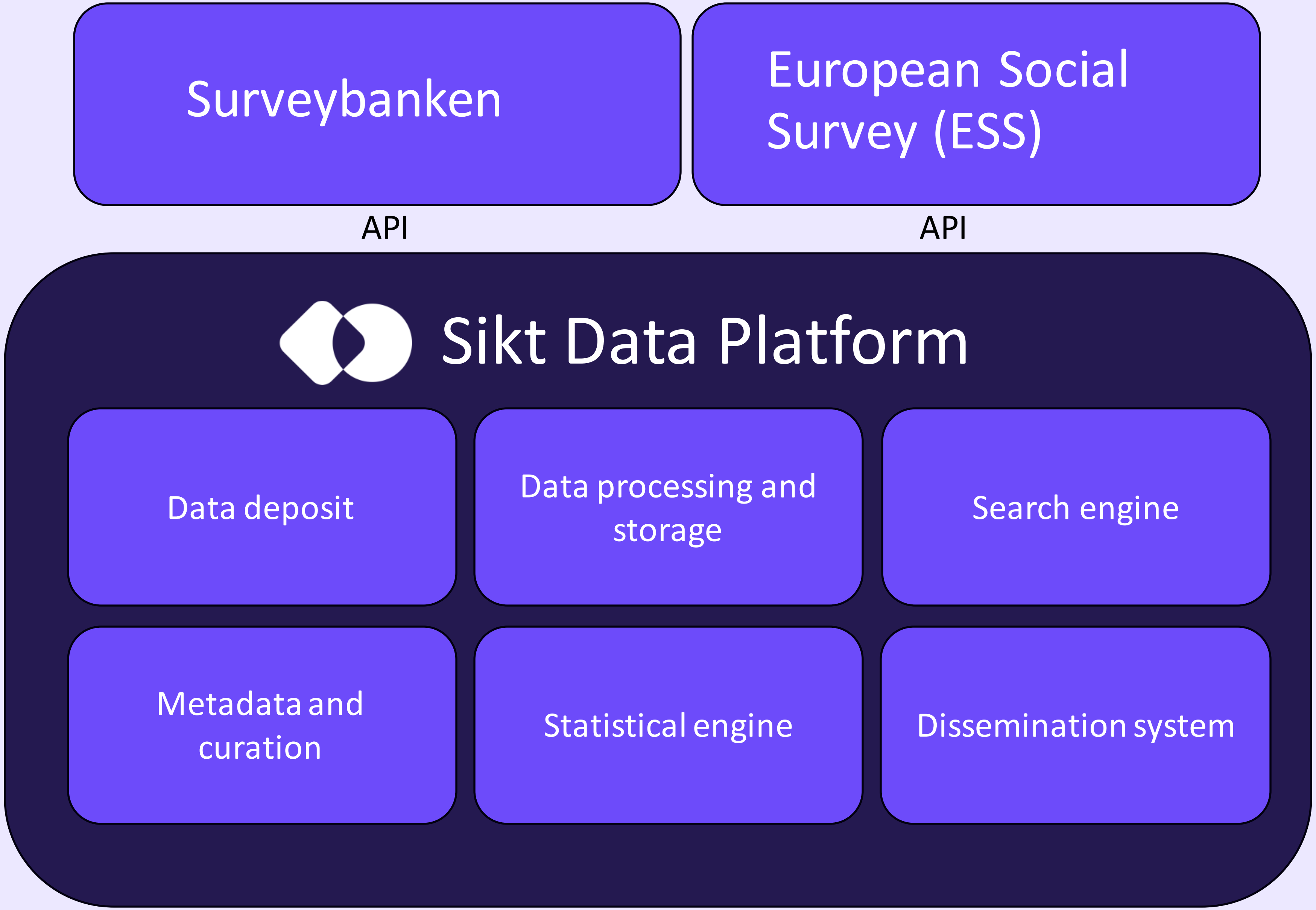
European Social Survey (ESS)

API

API

Sikt Data Platform

Metadata driven data infrastructure



Surveybanken

European Social Survey (ESS)

API

API

Sikt Data Platform

Data deposit

Data processing and storage

Search engine

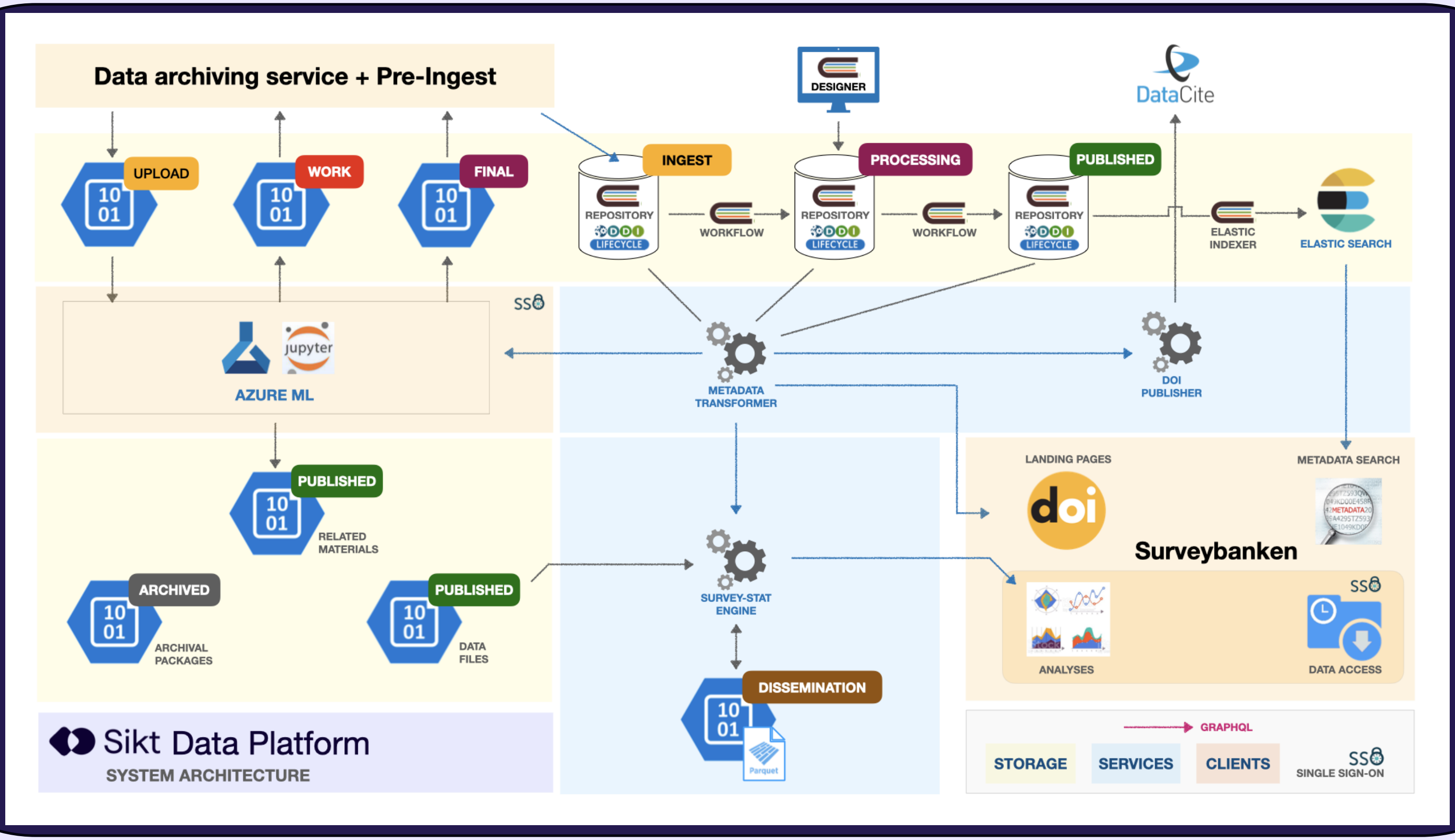
Metadata and curation

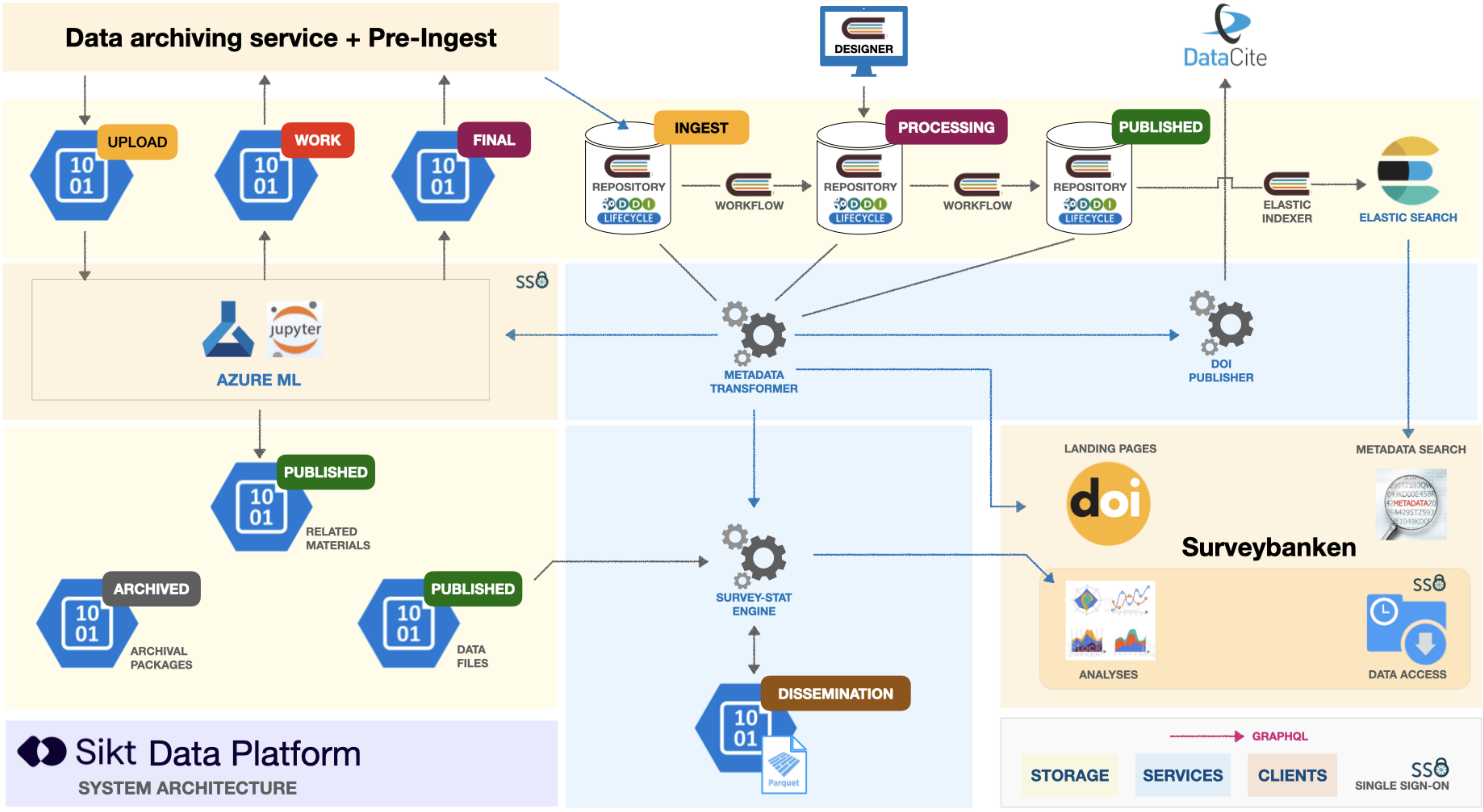
Statistical engine

Dissemination system

Surveybanken

European Social Survey (ESS)





Data archiving service + Pre-Ingest

UPLOAD

WORK

FINAL

INGEST

PROCESSING

PUBLISHED

REPOSITORY

LIFECYCLE

WORKFLOW

DESIGNER

DataCite

ELASTIC INDEXER

ELASTIC SEARCH

SSO

AZURE ML

jupyter

METADATA TRANSFORMER

DOI PUBLISHER

PUBLISHED

RELATED MATERIALS

ARCHIVED

ARCHIVAL PACKAGES

PUBLISHED

DATA FILES

SURVEY-STAT ENGINE

DISSEMINATION

Parquet

LANDING PAGES

METADATA SEARCH

Surveybanken

ANALYSES

DATA ACCESS

GRAPHQL

STORAGE

SERVICES

CLIENTS

SSO SINGLE SIGN-ON

Sikt Data Platform

SYSTEM ARCHITECTURE

Protocols adopted (1)

- DDI Lifecycle
  - encourages comprehensive description of data for discovery and analysis
  - facilitates machine-actionability and interoperability
- GraphQL API
  - designed to support strong types
  - offers flexibility for various clients consuming GraphQL APIs
  - encourage decoupling of the externally available API and it's internal implementation
- OIDC/OAuth (for login)
  - OAuth 2.0 is industry-standard protocol for authorization
  - OIDC is API-friendly
  - OAuth 2.0 capabilities are integrated in the OIDC protocol

Sikt

Protocols adopted (2)

- Terraform
  - let's you define infrastructure resources in human-readable configuration files that you can version, reuse, and share
  - provides a consistent workflow to safely and efficiently provision and manage your infrastructure throughout its lifecycle
- Apache Parquet
  - an open-source columnar data storage file format designed to support fast data access and analysis
  - compression is performed column by column and supports flexible compression options per data type
  - optimized for performance and supports data schema evolution
- JSON
  - an open text-based data-interchange format
  - easy for humans to read and write
  - machines can easily parse and generate it

Recommendations

AGILE APPROACH, CROSS FUNCTIONAL TEAMS

API FIRST

ENABLING METADATA DRIVEN SERVICES

BUILD DATA SOLUTIONS FOR DOWNLOAD AND ANALYSIS

CONNECT DATA AND METADATA

Thank you!

archana.bidargaddi@sikt.no
morten.jackobsen@sikt.no