Semantic Biobank Broker

Stefan Schulz, Markus Kreuzthaler

Vienna, March 14, 2018

C3med

BIOMARKER RESEARCH

Semantic Biobank Broker

Stefan Schulz, Markus Kreuzthaler

Vienna, March 14, 2018

Background: The FAIR guiding principles for data management

Manifesto for sustainable use of scientific research objects (data, workflows, algorithms) by humans and their digital agents

- **F – Findable** – Enriching datasets with metadata and annotation to support high quality content retrieval
- **A – Accessible** – Facilitating access to the data according to clear regulation regarding licenses of use
- **I – Interoperable** – Using machine-readable and internationally compatible standards for semantic annotations and metadata
- **R – Reusable** – Using exhaustive semantic annotations and metadata to reliably repurpose data, by preserving provenance, data production, and other contextual information.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3 (2016): 160018.

Problem Statement

- **Billions of biosamples available in biobanks across the world provide – in theory – enough material to support a broad range of biomedical research for the benefit of patients**
- **Current bottlenecks:**
  - How to find biosamples of a certain type of patients with a specific profile (Demographics, clinical history, main diagnosis, staging, grading, comorbidities, complications, biomarkers, therapies, survival, relapse)?
  - How to find out which biobank has samples of a certain type about a patient with a certain profile? Research on orphan diseases may require to contact many biobanks for few samples
  - How to get clinical / phenotypical data of interest together with biosamples?
  - How to attach old biosamples with recent clinical data of the same patient?
- **CBmed take: to propose and discuss feasibility of a "Biobank broker", following the FAIR principles**

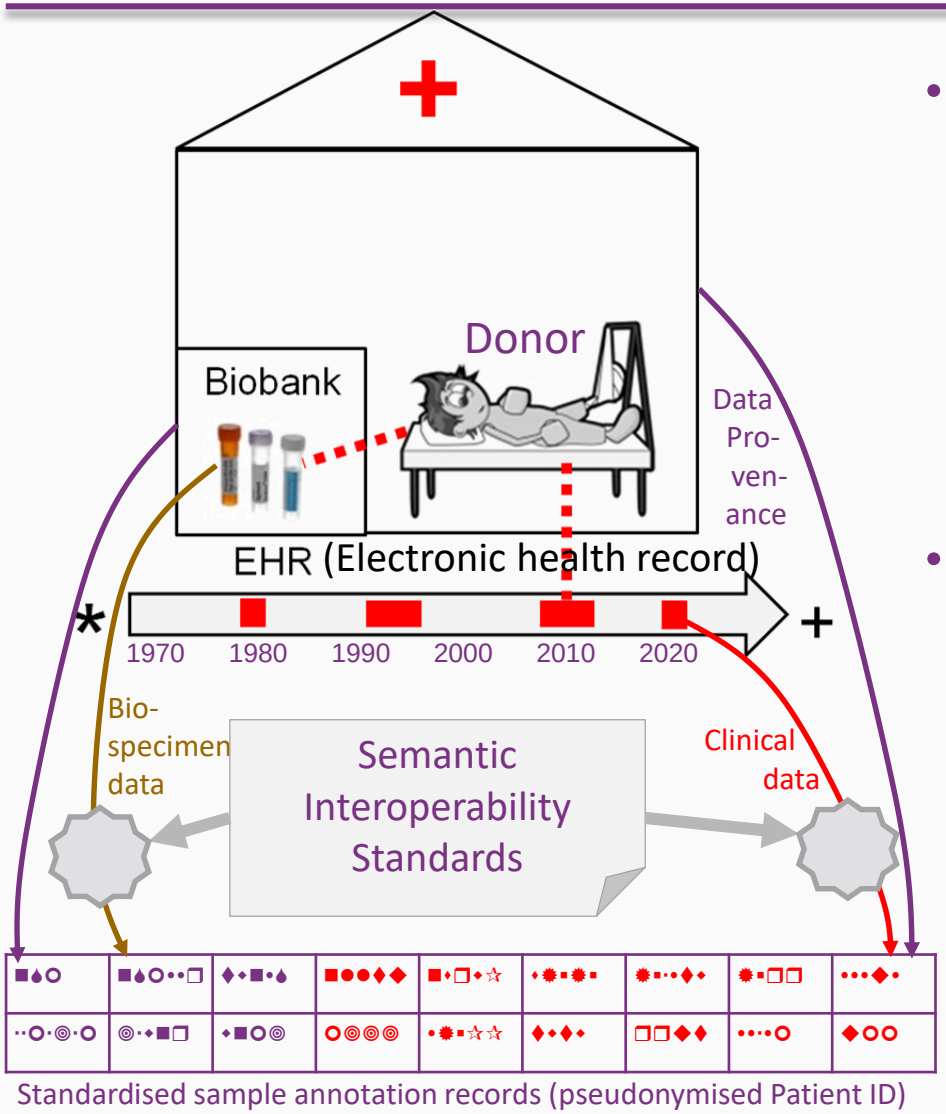
Abstraction of the problem

- **Characteristics**
  - Heterogeneous resources
  - Distributed across the world
  - Requested for multiple purposes
- **Functional desiderata**
  - Effective retrieval
  - Across many axes
  - Using a common, powerful query language
  - Regulate access
- **Technical requirements**
  - Orchestration of resource-level and meta-level search
  - Semantic annotation standards
  - Quality, provenance

Analogous scenarios

- **Literature search**
  - Heterogeneous, distributed resources (books, journals)
  - Cross-resource search by literature databases
  - Standardised annotation vocabulary (e.g. MeSH in MEDLINE)
  - Comprehensive query engines (Pubmed, OVID,…)
- **Special purpose federated search engines**
  - Flights, car rental, travel (Google Flights, TripAdvisor, Trivago)
  - Central search request passed on to numerous search engines
  - Results fetched from search engines, aligned, fused and ranked

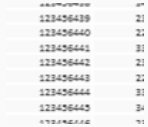
Biobank-related data

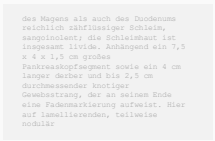


- Sample related information:
  - Type
  - Quality
  - Time
  - Storage information
  - Physical location
  - Lab data, genotype,…
- Donor related information:
  - Demographic data
  - Phenotype data
  - Time indexed clinical data (patient record extracts)
  - Increment of clinical information after sample is taken

Semantic standardization

Structured data:
- Administrative codes
- Lab
- Registries





Textual data:
- Finding reports
- Problem lists
- Discharge summaries

Electronic health record

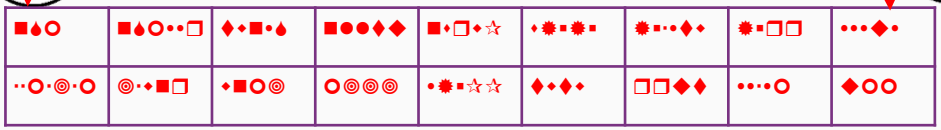
1970 1980 1990 2000 2010 2020

Annotation by
Coding experts
(local or remote service)

Information standards
(HI7-CDA, openEHR,…)
Terminology standards
(SNOMED CT, LOINC, ICD ,…)

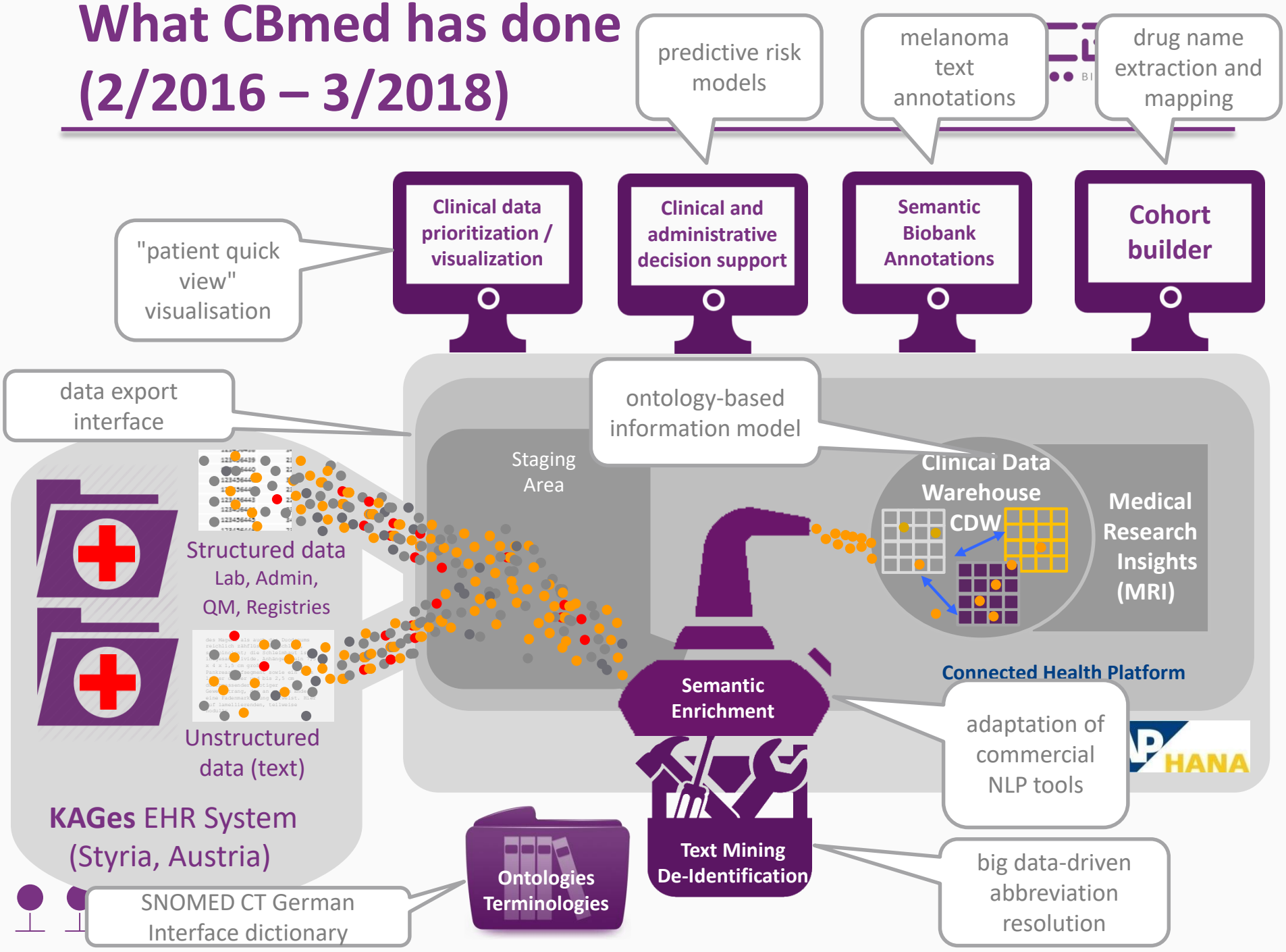
Local
dictionaries

Annotation / Information
Extraction using NLP
(local or trusted cloud service)



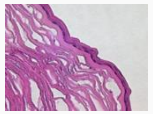
Standardised annotation record

What CBmed has done (2/2016 – 3/2018)



Functionality of a Biobank broker compared to MEDLINE search

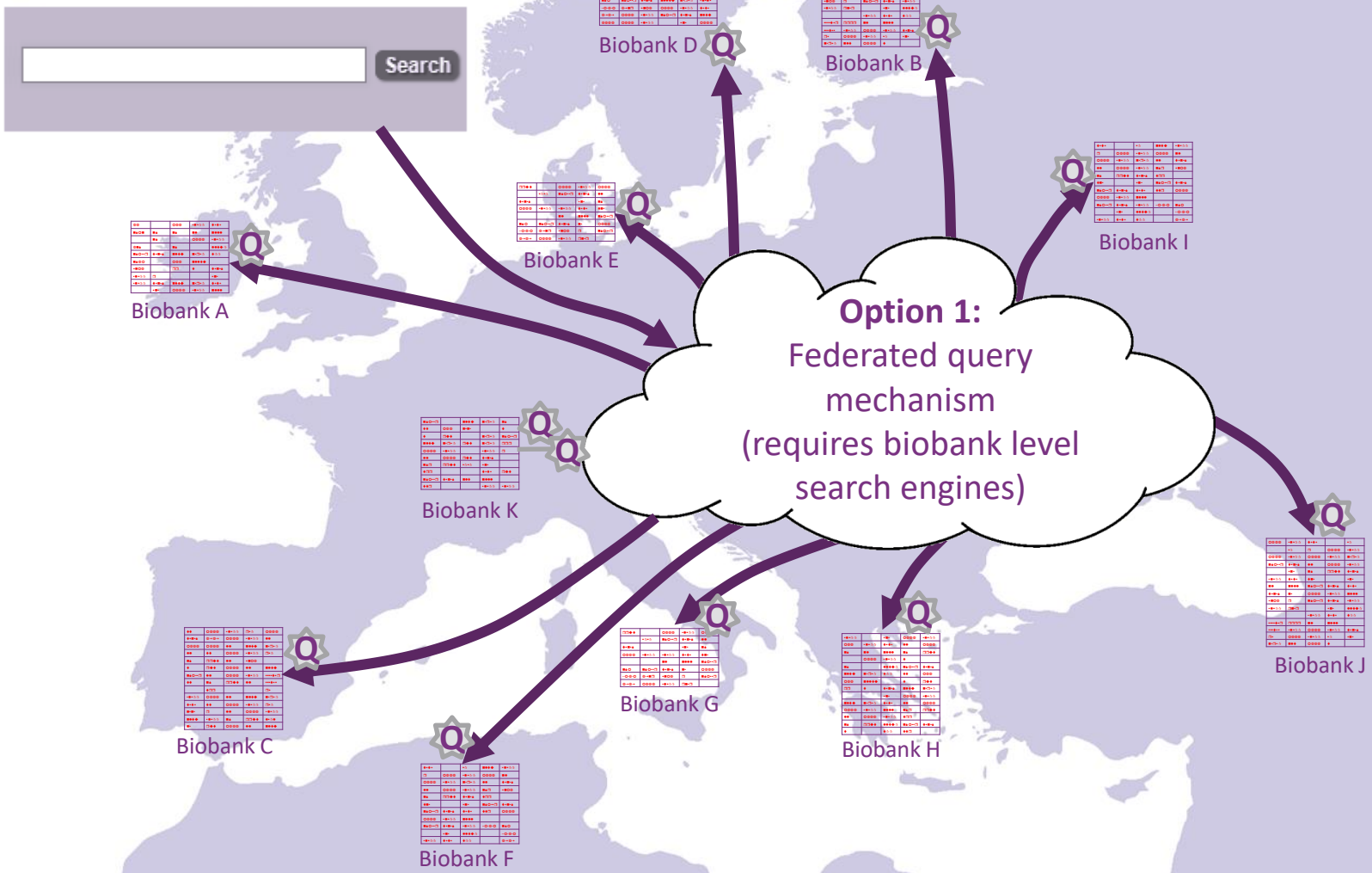
Biobank "Broker"



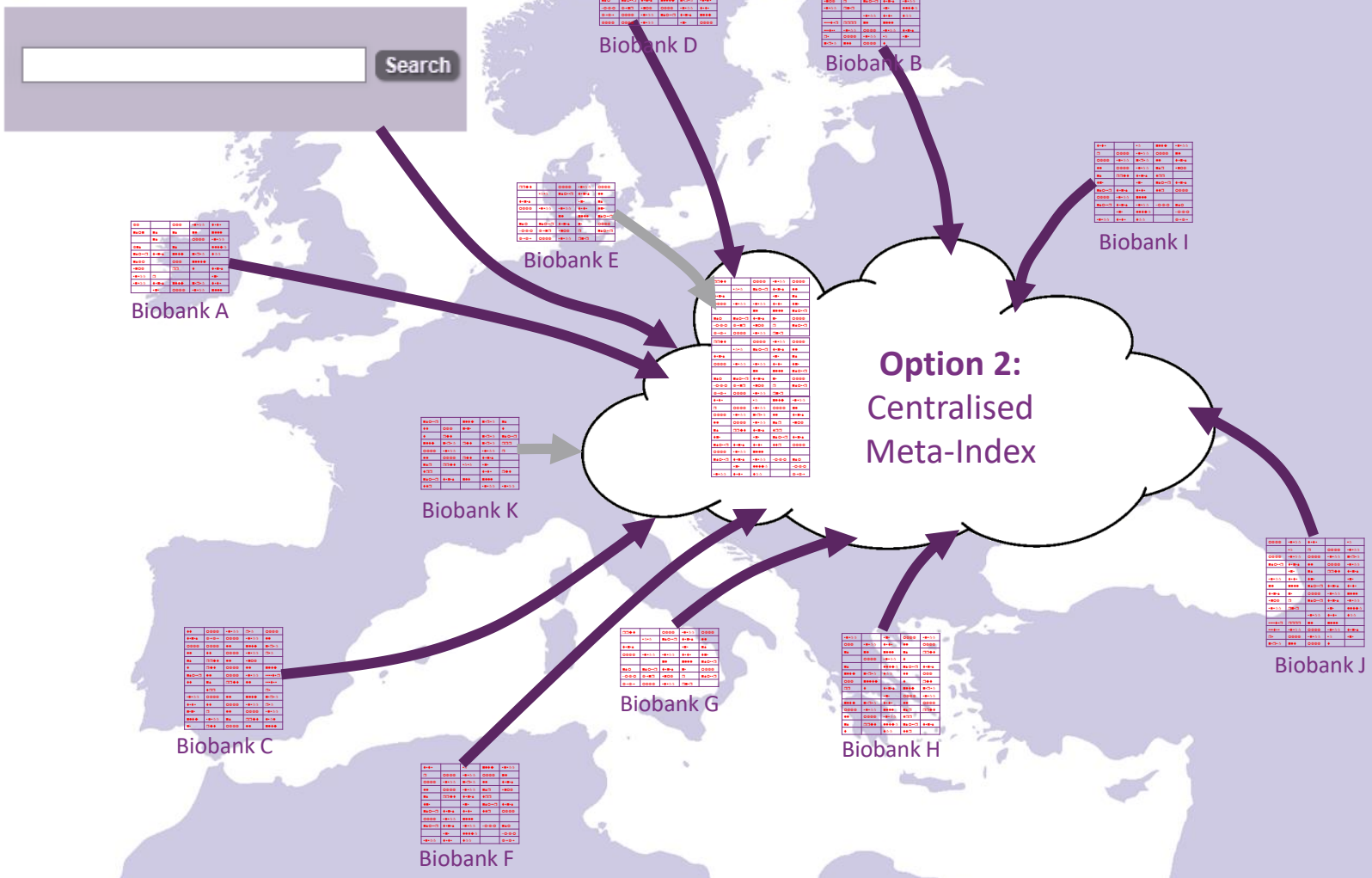

- Global bibliographic database
- Resources: publications from different publishers
- Annotations:
  - Bibliographic data
  - Abstract
  - Semantic representation (MeSH) of paper content
- Full resource (papers) access:
  - Needed by many users
  - Restrictions apply (paywall)

- Global biobank sample database
- Resources: biological specimens (blood, tissue,…)
- Annotations:
  - Sample information
  - Semantic representation of selected patient related information (Information models / ontologies)
- Full resource (EHRs) access:
  - Not needed by most users
  - Restriction apply (privacy)

Biobank broker



Biobank broker



Challenges / Open issues (I)

- **Standardised information templates**
  - Templates tailored to different kinds of samples / diagnoses
  - Consensus regarding content scoping, terminologies and information models, context parameters
- **Federated queries versus centralised repository**
  - Design and maintenance of multiple interfaces
  - Low runtime performance issues when relaying queries to multiple local query engines
  - Centralized meta-index timeliness of data not optimal, transfer of local data to central hub may contradict regulations (even if de-identified)

Challenges / Open issues (II)

- **Interfaces to clinical data**
  - Multiple natural languages, interface dictionaries to be created
  - Interfaces between raw data (plain text?) and NLP engine
  - Training data – specific to language, documentation systems, clinical specialties
- **Data processing platform at biobanks and clinical centres**
  - Annotation: Human, machine (NLP), hybrid
  - Annotation guidelines; training of annotators
  - Stand-alone application or secure cloud service?
  - Integration with local data warehouse solutions

Challenges / Open issues (III)

- **NLP services**
  - Depend on existing language resources and training data. Local vocabularies need frequent updates and alignment with semantic standards
  - For important information needs (e.g. smoking status), costly crafting of dedicated information extraction tools
- **Data quality**
  - Quality issues at source, e.g. ICD codes for administrative coding
  - Clinical documents: misspellings, telegram style, abbreviations, contexts, time markers
  - Tabular data: unclear contexts, local value sets
  - Constant quality management required

Challenges / Open issues (IV)

- **Ethics**
  - Different regulation across institutions
  - Pseudonymisation / re-identification
- **Governance**
  - Access regulations
  - Update policies
  - Sustainability / business model
- **Usability**
  - Graphical user interface
  - Terminology support

Stakeholders

- **Stakeholders in Graz**
  - CBmed: Rapidly growing global player in biomarker research
    - Closely connected with Medical University of Graz
    - Synergies with publically funded and industry-sponsored activities
  - Medical University of Graz:
    - Hosts largest European biobank
    - Reference in biomedical semantics and terminology research
    - Experience in clinical querying
    - Unique co-operation with one of the largest Austrian hospital networks (KAGes), pioneer in large-scale clinical warehousing
    - Successful partnering with SAP (clinical data warehousing and querying)
  - BBMRI-ERIC: Europe-wide biobank hub
    - optimally connected with numerous European biobanks
- **Stakeholders across Europe**
  - SAP: warehousing, data management and cloud technology provider
  - Biobanks connected with BBMRI-ERIC
  - Pharma industry as potential clients for future semantic biobank services

Skills matrix

	CBmed	Meduni Graz	BBMRI-ERIC	SAP	Biobanks	Pharma
Annotation templates			x		x	
Use cases	x		x		x	x
Terminology standards		x				
Information modelling		x	x	x		
Data warehousing				x		
Cloud technology				x		
Language technology	x	x		x		
EHR interfaces		x			x	
Human annotation					x	
Machine annotation	x	x				
Clinical big data	x	x				
Legal / ethical issues	x	x	x			
Business modelling	x			x		x

Proposed actions

Year	Phase	Action
1	Requirement analysis	▪ Analyse roles and needs of partners, particularly potential clients ▪ Create inventory of tools resources needed ▪ Analyse existing work, discuss reuse of related materials, drafts, models etc. ▪ Harmonise with BBMRI-ERIC goals and policies ▪ Select preferred architecture ▪ Identify legal and ethical challenges
2	First prototype	▪ Collect use cases and test queries, make selection ▪ Define appropriate information templates ▪ Build / adapt terminological resources (language-dependent) ▪ Develop tooling for manual and machine annotation, dependent on partner ▪ Implement cloud-based broker architecture ▪ Build demonstrator and assess
3	Refined prototype	▪ Additional use cases ▪ Extending and improving tools and resources ▪ Implement cloud-based solution for annotation at one site ▪ Enhanced demonstrator, evaluation, dissemination of results
4	Consolidation	▪ Request of external funding ▪ Stakeholder analysis / value proposition ▪ Business model / sustainability strategy ▪ Definition of next project

References

- Andrade, A. Q., Kreuzthaler, M., Hastings, J., Krestyaninova, M., & Schulz, S. (2012). Requirements for semantic biobanks. Stud Health Technol Inform, 180(1), 569-73.
- Müller, H., Reihs, R., Zatloukal, K., Jeanquartier, F., Merino-Martinez, R., van Enckevort, D., ... & Holzinger, A. (2015). State-of-the-art and future challenges in the integration of biobank catalogues. In Smart Health (pp. 261-273). Springer, Cham.
- Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., ... & Van Der Lei, J. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics, 216, 574.
- Brochhausen M1, Fransson MN, Kanaskar NV, Eriksson M, Merino-Martinez R, Hall RA, Norlin L, Kjellqvist S, Hortlund M, Topaloglu U, Hogan WR, Litton JE. Developing a semantically rich ontology for the biobank-administration domain. J Biomed Semantics. 2013 Oct 8;4(1):23. doi: 10.1186/2041-1480-4-23.
- Ontology for Biobanking: https://bioportal.bioontology.org/ontologies/OBIB
- SNOMED CT: https://www.snomed.org/snomed-ct
- LOINC: https://loinc.org
- HL7: http://www.hl7.org