Data First –
Information Procurement based on FAIR Principles

Martin Romacker
Data and Information Architect
Pharma Research and Early Development Informatics (pREDi)
Roche Innovation Center Basel

Prisme Forum, 15th November 2018, Chicago

PRISME Forum
Pharmaceutical R&D Information Systems Management Executive Forum

Roche pRED

An Episode

Data Quality and Data Curation

An Episode

- Pistoia Alliance Conference Mar 2018 London (approx 180 participants)
  Pistoias Alliance: global, not-for-profit alliance of life science companies, vendors, publishers, and academic groups that work together to lower barriers to innovation in R&D
- 4 Breakout Sessions: Lab of the Future, Artificial Intelligence, Real World Data and Data FAIRification
- Slightly more than 10 people showed up at RWD/ FAIR breakout
- Data FAIRification of relevance but difficult to promote
- AI breakout came back with a ranked list of 6 high priority activities:
  First and most important item – high quality well-curated data !!!

→ Data Quality and Interoperability are strangely neglected topics

Management of Corporate Data Assets

Assessment

- Competence Center for Corporate Data Quality (University of St Gallen, CH)

Institute of Information Management

University of St.Gallen

CDQ

Corporate Data Quality

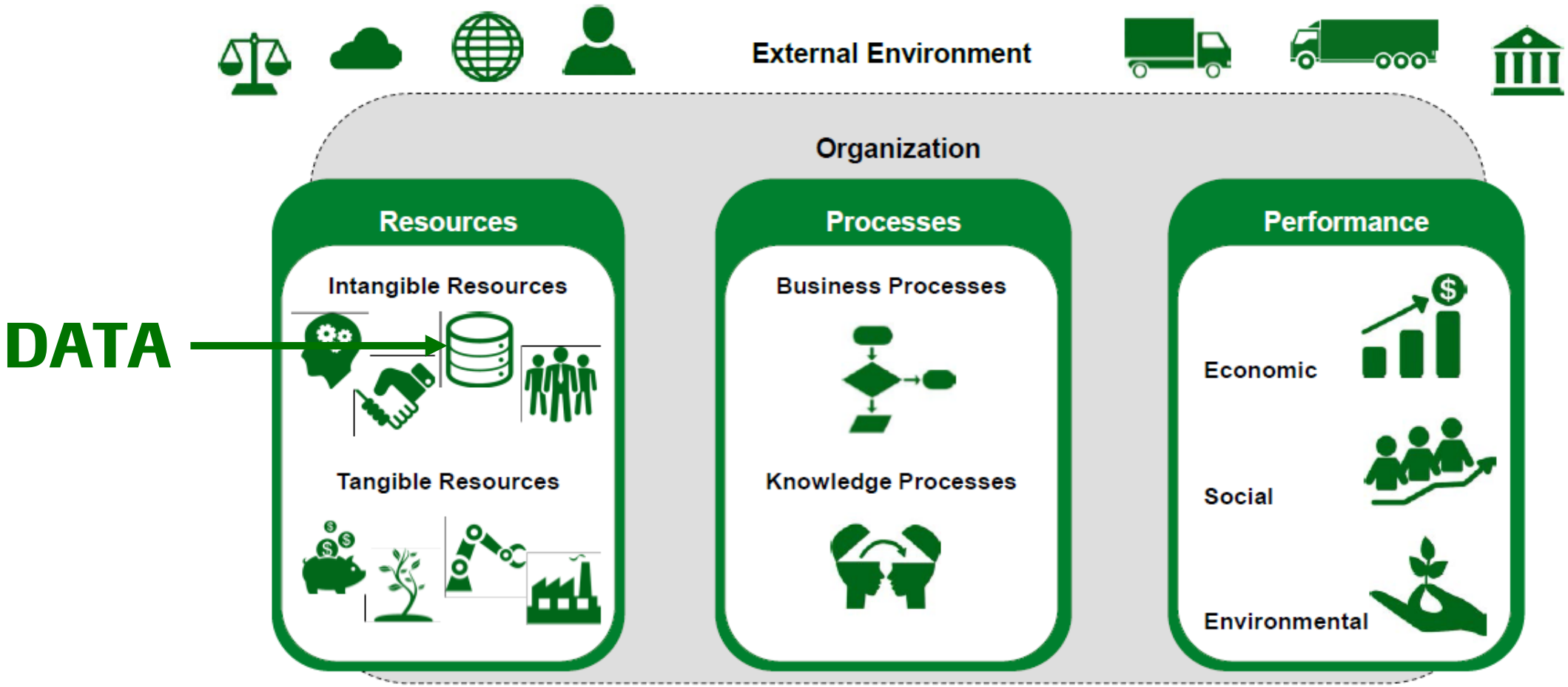
- Andreas Zechmann PhD Student

Institute of Accounting, Control and Auditing
Chair of Controlling / Performance Management
Prof. Dr. Klaus Möller

Data asset performance management
An approach towards data asset valuation

Management of Corporate Data Assets

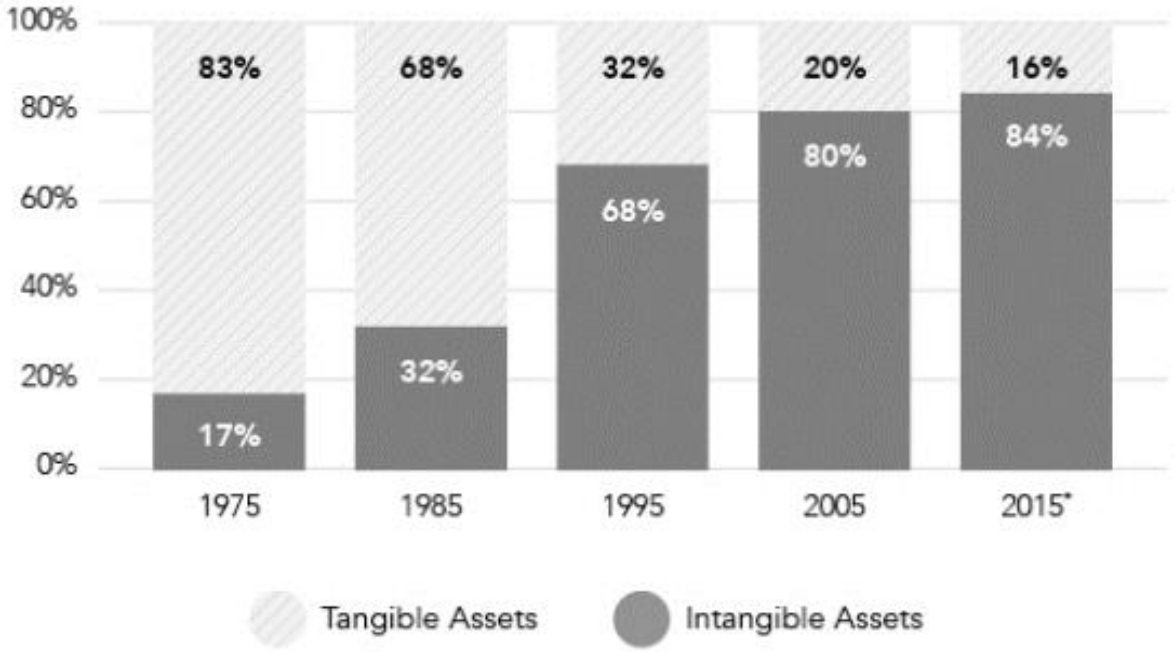
Intangible Resources



Management of Corporate Data Assets

Insufficient Management of Intangible Resources

Components of the S&P 500 Market Value[1]



How CEOs recognize data as an asset[2]

Prof. Holger Kohl *(Fraunhofer competence center for knowledge management)*:

»Companies increasingly recognize the **importance** of **intangible resources** […], but they still **manage** these important **factors** for **business success insufficiently**«[3]

Management of Corporate Data Assets

Value Measurements

	Methods for the valuation of intangible assets[1]		
	Market approach	**Cost approach**	**Income approach**
Key question:	If I were to sell my data, how much would I get for it?	How much would it cost me to reproduce my current data?	What is the value that my data generates in the business processes?
Concept:	Based on **market prices** or **multiples**	**Reproduction** or **replacement cost**	**Present value** of **earnings** attributable to the asset or costs avoided
Applicability:	In many cases, markets and market prices do not exist. **Often not suitable.**	Method considers reproduction costs and subtracts adjustment charges for quality and usage. **Suitable.**	Method considers risks and income derived from data. Value assessment based on estimations. **Suitable.**

Shift Happens – Towards a Data Driven Industry

Foundational Change: Perception of Value of Data

Pharma Times online, 09. November 2016

British AI group licenses Janssen drug candidates

British artificial intelligence group BenevolentAI has signed an exclusive license with Johnson & Johnson group Janssen, picking up rights to a series of its novel clinical stage drug candidates. Under the deal, BenevolentAI has acquired a li
select number of candidates and their extensive related portfolio of patents, after concluding that there is "strong promise" to develop them into new medicines for hard to treat diseases using its artificial intelligence technology.

Read more | Open in app

Diese Verschiebung haben die Unternehmen laut PwC mit gutem Grund vorgenommen: Diejenigen Firmen, die im Branchenvergleich schneller als der Durchschnitt gewachsen seien, hätten 2015 im Schnitt 25% mehr Geld für Software-Entwicklung ausgegeben als die Unternehmen, deren Umsatz sich unterdurchschnittlich entwickelt habe, heisst es in der Studie.

IT as Key Enabler

ROCHE IST AUF RANG 7 ABGERUTSCHT

Gleichzeitig sind Unternehmen, die in der Öffentlichkeit als besonders innovativ gelten, nicht unbedingt diejenigen, die auch tatsächlich am meisten in F&E investieren. So führt wie im Vorjahr der Automobilkonzern VW die Liste mit Ausgaben von 13,2 Mrd USD an. Samsung folgt auf Platz zwei, gefolgt von Amazon.

Apple dagegen, das als innovativstes Unternehmen gilt, landet mit Ausgaben von 8,1 Mrd USD lediglich auf Platz 18. Alphabet, die Muttergesellschaft von Google, kommt mit 12,3 Mrd immerhin auf Platz vier.

Pharma 4.0

Reimagining Novartis as a 'medicines and data science' company

Vas Narasimhan on LinkedIn

January 11, 2018

Machen Google und Co. schon bald Jagd auf Roche und Novartis?

Mit der milliardenschweren Übernahme von Whole Foods durch Amazon hat das Detailhandelssterben jenseits des Atlantiks einen traurigen Höhepunkt erreicht. Reihenweise sind kleinere Ladeninhaber gezwungen, ihre Geschäftstätigkeit aufzugeben. Gegen den mächtigen Versandhändler scheint kein Kraut gewachsen.

Der Vorstoss der Amerikaner ins Geschäft mit Nahrungsmitteln sollte auch anderen Wirtschaftszweigen eine Warnung sein. Denn immer öfter nutzen Tech-Giganten wie Amazon, Googleund Co. ihre Milliarden von Dollar, um sich neue Märkte zu erschliessen.

SPIEGEL: Muss Roche mehr wie Google werden, sich zum Datenkonzern wandeln?

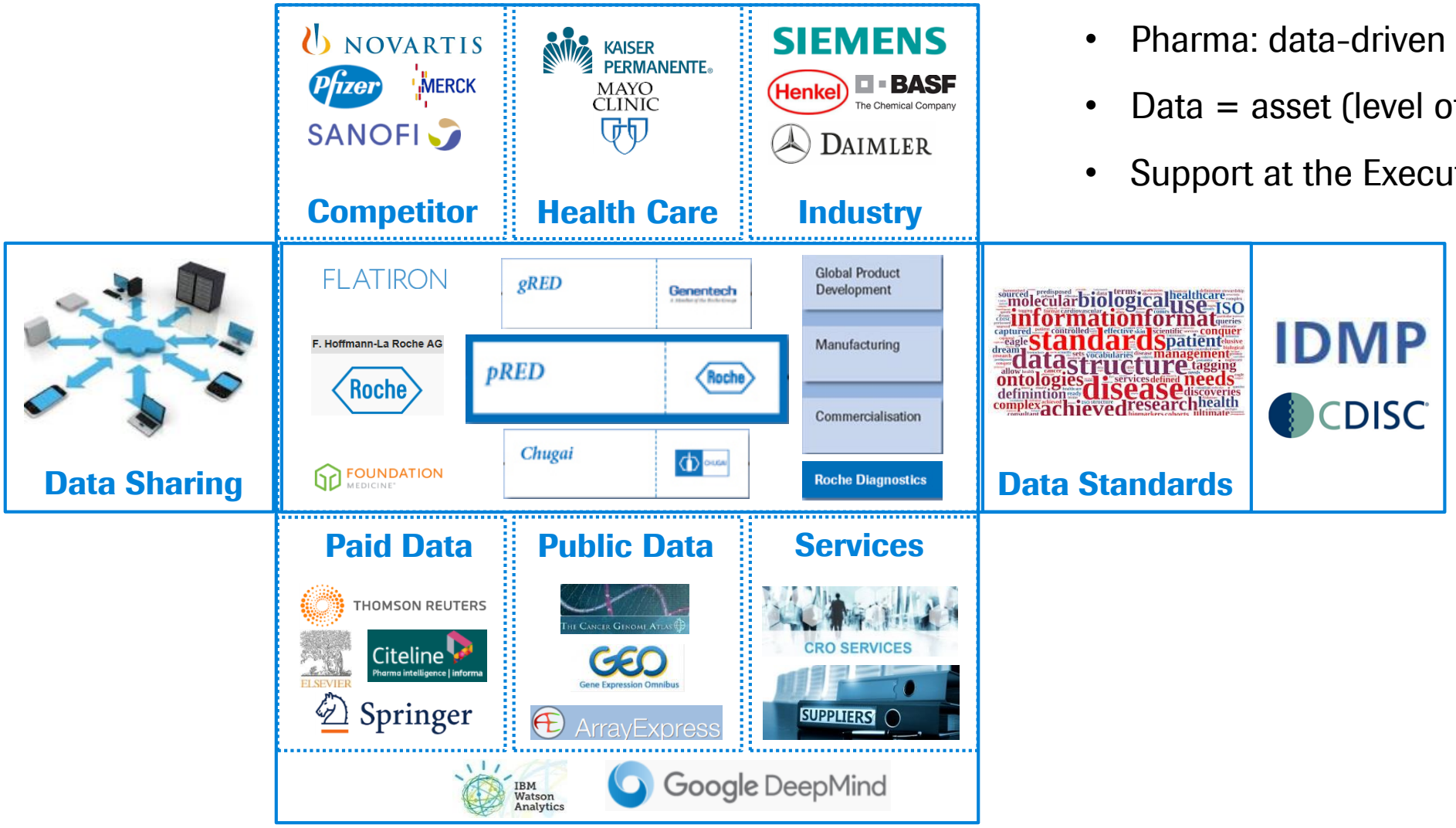
Franz: Absolut. Google muss erst die Pharmaerfahrung aufbauen – und wir müssen die Digitalisierung für uns nutzen. Mit den Produkten unserer Diagnostiksparte generieren wir im Jahr 15 Milliarden Tests, also Datenpunkte. Wir haben einen riesigen Datenschatz. Aber wir fangen erst an, ihn zu nutzen.

«99% of the innovation is happening outside our companies»
(Severin Schwan, CEO F. Hoffmann-La Roche)

Data are no longer a by-product of business processes but move to the center of the strategy

Urgency for a new Information Landscape

Business Rationale: Dissolution of Boundaries

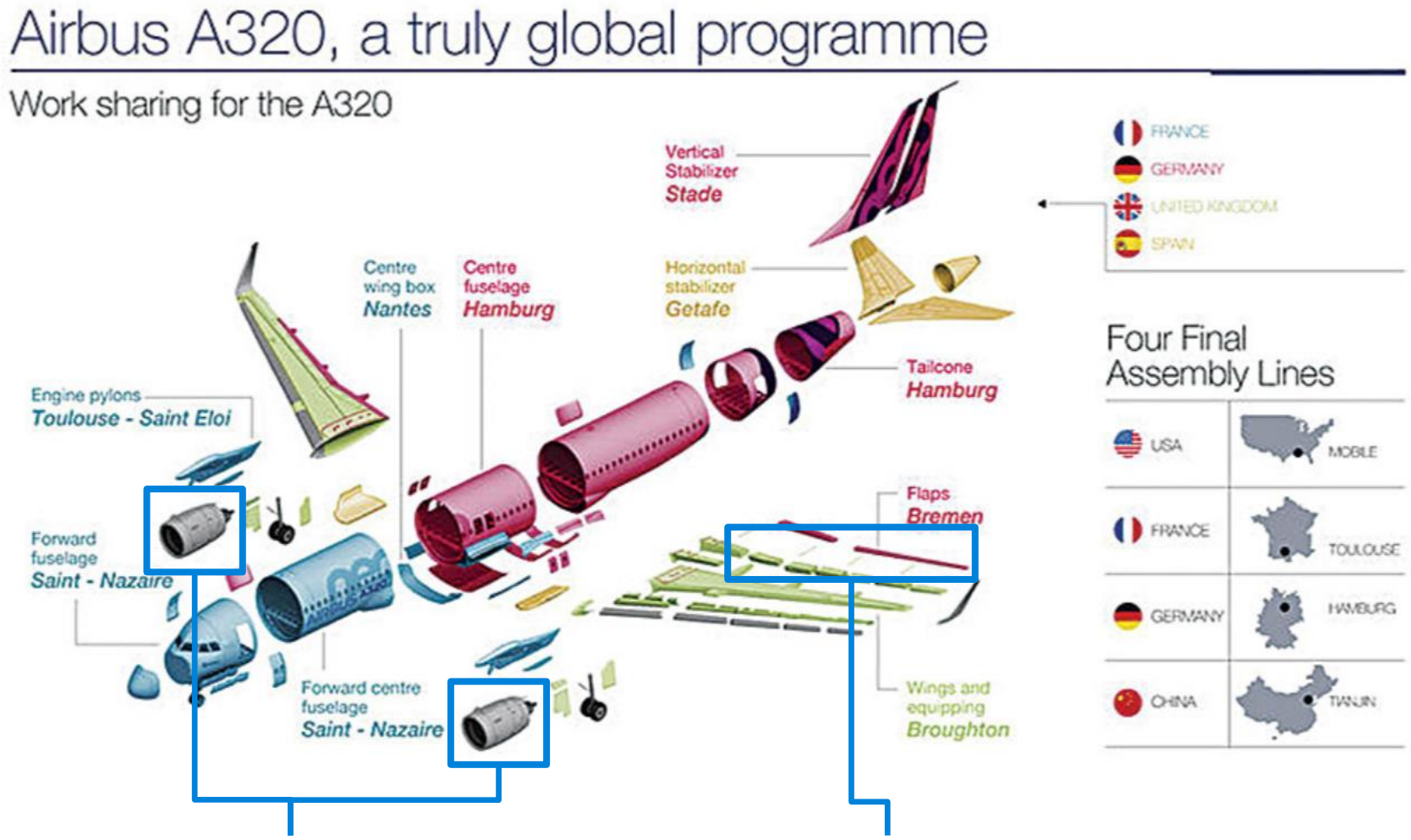


- Pharma: data-driven industry
- Data = asset (level of investment)
- Support at the Executive Level

Translational Medicine, PHC and RWD integration require a global cross-industry data management strategy

Manufacturing

Components & C-Parts Procurement



- highly complex product
- highly distributed process
- vertical integration (CMO)
- assembly in different global geographic regions

Are data-driven information assembly processes different?

Contract Manufacturing Organization

Technical Connections/ Interfaces

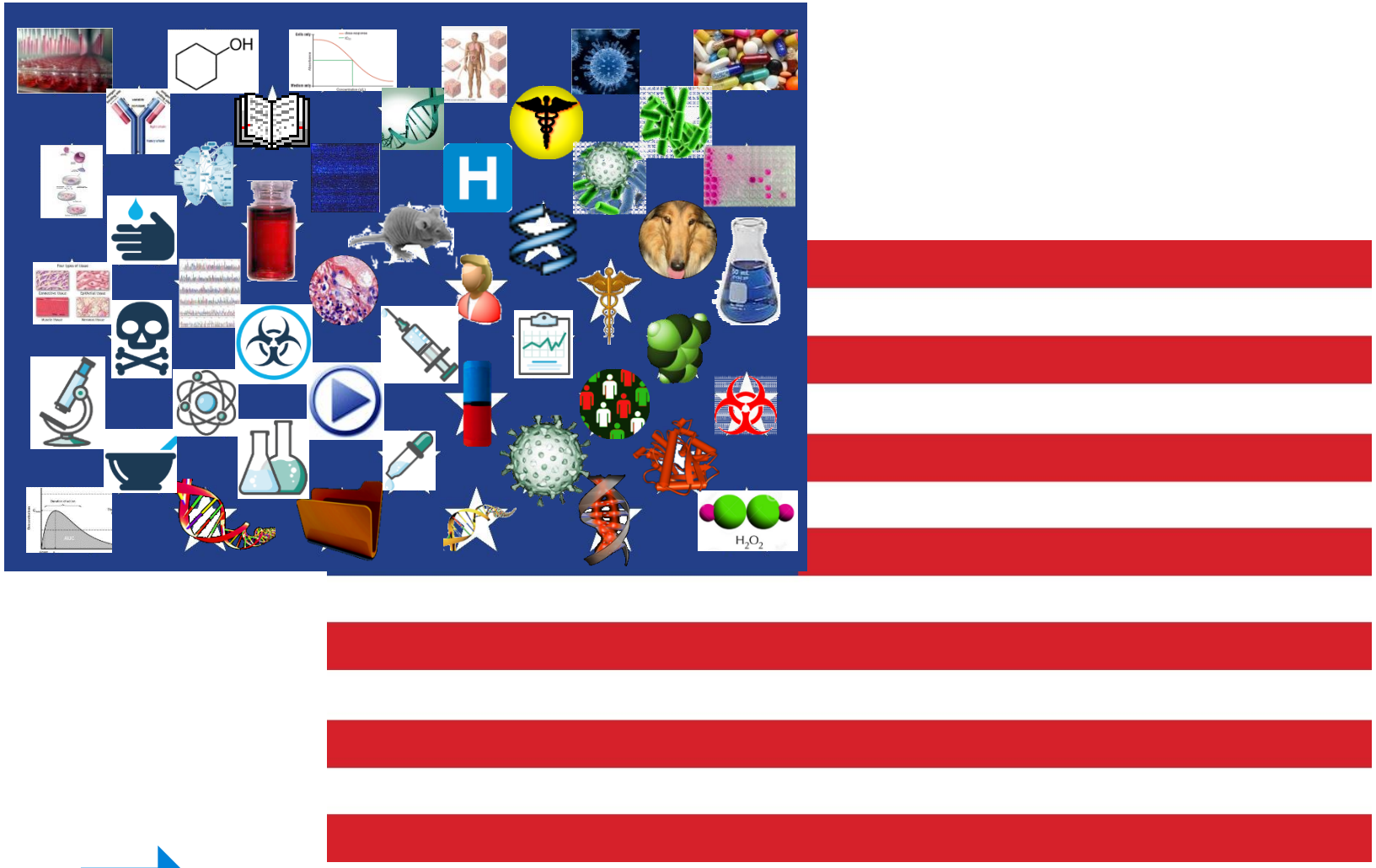
Information Procurement

Data-Driven and Knowledge-Based Pharma R&D

- Information Procurement:
  the effective and efficient process of creating, acquiring and integrating standardized information objects and data types into data-driven R&D activities
  - creation: an in-house activity which will result in a new data asset
  - acquisition: an internalization of a new data asset created by an external oganization
  - integration: assembly of internal and external datas asset into larger meaningful assets
  - information object: primary building block for information with definitions (eg indication, adverse event, gene, active ingredient) and their instantiations (eg COPD, lung cancer)
  - data type: information objects extended by metadata (eg cell line, assay, drug product) and their instantiations

Data First !

Knowledge Space in Pharma R&D



Huge universe of scientific and business objects to be represented

Data Quality, Data Curation & Artificial Intelligence

AI assisted Data Curation

- AI will not replace data curation – it might increase the productivity (urgently needed)

Experts say IBM Watson's flaws are rooted in data collection and interoperability, *fiercehealthcare.com*
Experts again called into question the abilities of IBM's premier artificial intelligence system Watson to revolutionize cancer care. Watson has two core issues that other systems have as well: interoperability and data collection. The input/output ratio of Watson is too low, while the sheer amount of data required to feed the system's deep learning network is another obstacle. Some also suspect Watson's results to be biased due to its main source of diagnostic information coming from a group of experts at Memorial Sloan Kettering.

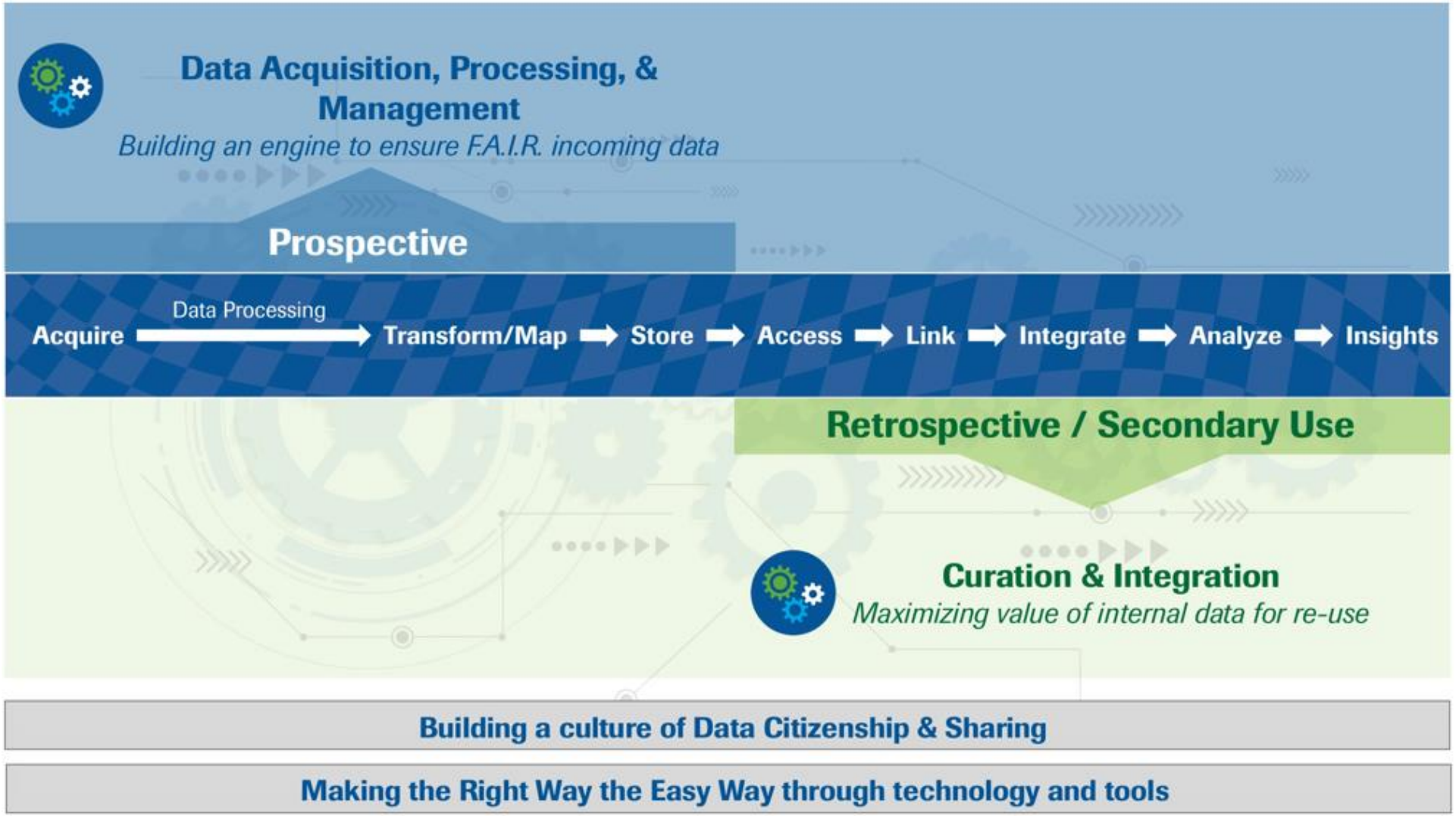
If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Fully Integrated Data Management Strategy build on FAIR

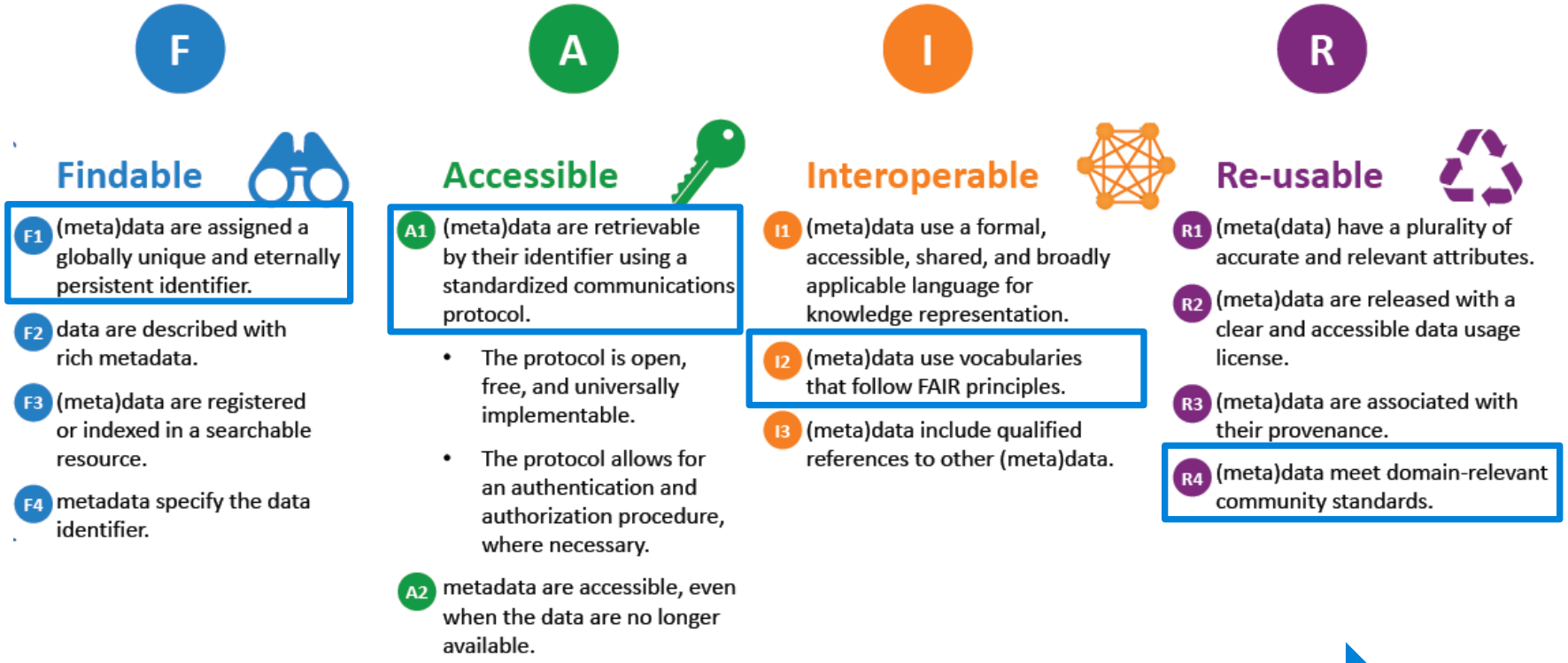
Semantic Data Management



Slide by Kimberly Barnholt, Program Manager, PD Biometrics

Making Data FAIR

Implications



- Data FAIRification: foundational change in the organization affecting
  - data and knowledge engineering
  - enterprise architecture (eg Roche Data Commons)
  - data and information flow
  - corporate culture

5-10 year effort

Source: https://www.dtls.nl/fair-data/fair-principles-explained/

Data FAIRification

Impact on Business, IT & Science

Information Procurement

- Standardization of scientific information:
  knowledge assembly line

Information Architecture

- Information centric organization of knowledge
- Clear definition of information objects across functions and divisions
- Metadata Registry
- Terminology Service
- Alignment of Standards
- Definition of shared Minimal Models
- Single Point of Truth/ System of Record
- Common language within organization, industry, vendors & academia

Productivity/ Insights

- Reduction of efforts for data integration and sharing (cost avoidance)
- Reduction of time to collect, normalize & analyze data (time to market)
- Significantly improved data quality (less variety, increased veracity)
- Better science due to less fragmented, easier to access and faster to integrate data
  (value generation)

Governance

- Definition of data sharing principles (black list vs white list) – but GDPR etc.
- Data ownership
- Process ownership
- Make the right way the easy way: data capture/ data curation

Data FAIRification

IT Capabilites Supporting Business Processes

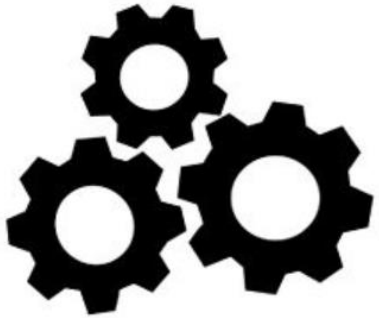
Findable

Terminology
Identifer Resolution
Data Catalog
Semantic Search

Accessible

Standard API (REST)
Resources – URI
Metadata (eg DCAT)
Data Sharing Policies

Interoperable



Terminology
Metadata Registry
Business Glossary
Data Standards (eg CDISC)

Reusable

Data Quality KPIs
Minimal Models
Data Validation (FAIR metrics)
Data Governance

Take knowledge to the point of usage

Data FAIRification

Data Quality KPIs (reusability, reproducibility)

KPI		
Completeness	all mandatory data elements are present & captured	✡ ✡
Correctness	all data elements are correct	✡ ✡
Conformity	all data elements conform to a defined standard	✡ ✡
Consistency	all data elements are consistent (meaningful record)	✡
Coherence	all data elements are coherently applied across applications	✡
Currency	all data elements are up to date (life cycle)	✡

✡ Data Curation – intellectual process of information recovery/ validation

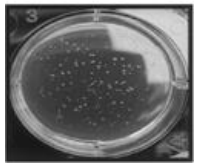
✡ Data Standards (Terminology, Minimal Models & Metadata Registry)

- ISO 25012:2008 Dimensions of Data Quality
- ISO 8000 Series for Data Quality (eg 8000:100 Terminologies, 8000:120 Provenance)
- ISO 38505: Governance of Data

Information Procurement

(a negative) Example

NCI-H522



Cell Line Registry 1

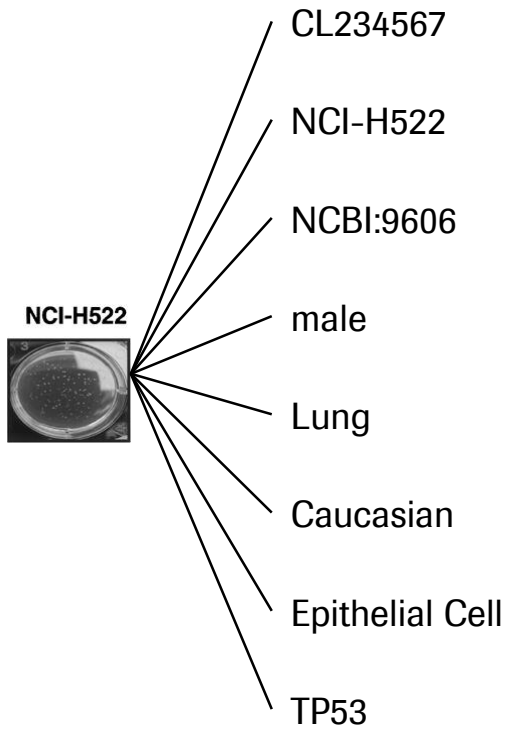
Cell Line ID: CL23456
name: NCI-H522
hasSpecies: NCBI:9606
gender: male
tissueOfOrigin: Lung
morphology: Epithelial Cell
ethnicity: Caucasian
mutation: TP53

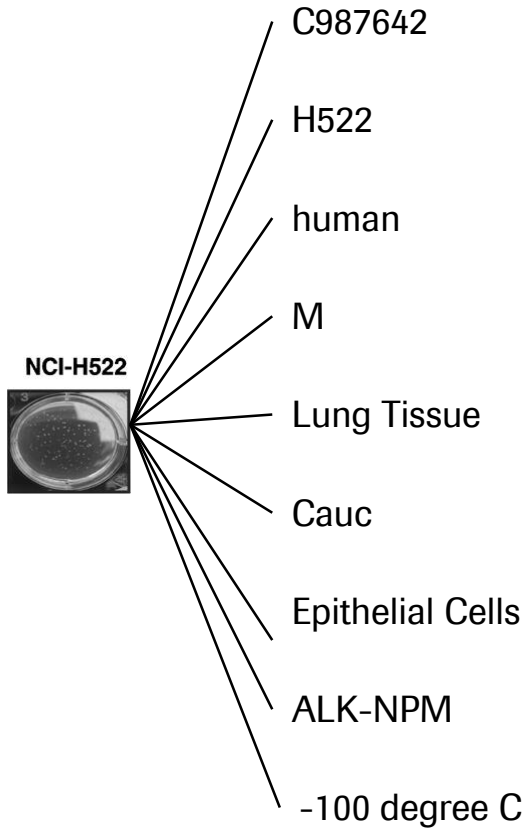
Cell Line Registry 2

ID: C987642
label: H522
species: human
sex: M
tissue type: Lung Tissue
cell type: Epithelial Cells
race: Cauc
gene fusion: ALK-NPM
freeze condition: -100 degree C

Semantic Integration

Initial Setup

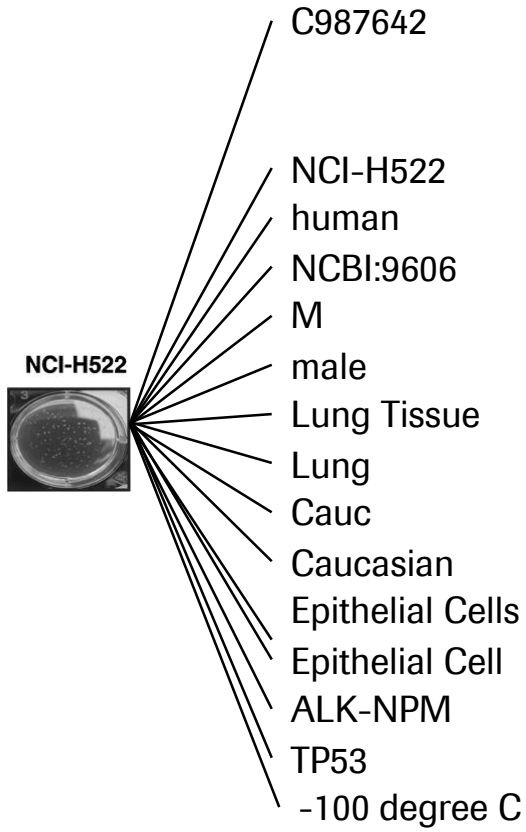




Cell Line ID	ID	Name	Label	Has Species	Species	Gender	Sex	Tissue of origin	Tissue	Ethnicity	Race	Morphology	Cell type	Mutation	Gene Fusion	Freeze Condition
CL234567		NCI-H22		NCBI:9606		Male		Lung		Caucasian		Epithelial Cell		TP53		
	C987642		H522		human		M		Lung Tissue		Cauc		Epithelial Cells		ALK-NPM	-100 degree C

Semantic Integration

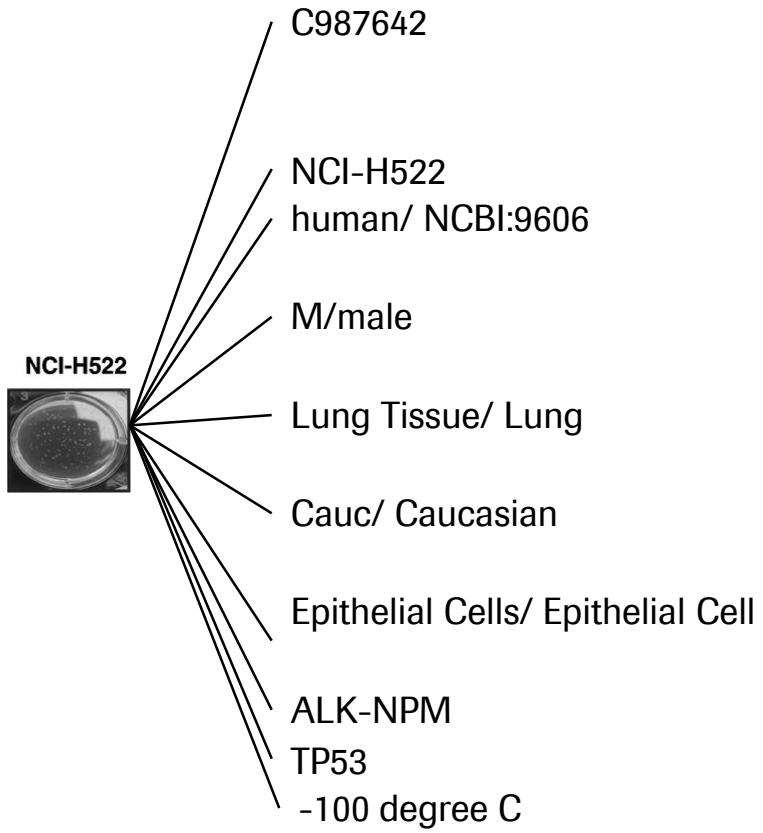
Identity Resolution (Glossary, Identifier Service)



Cell Line ID	Preferred label	Has Species	Species	Gender	Sex	Tissue of origin	Tissue	Ethnicity	Race	Morphology	Cell type	Mutation	Gene Fusion	Freeze Condition
CVCL1567	NCI-H22	NCBI:9606		Male		Lung		Caucasian		Epithelial Cell		TP53		
CVCL1567	NCI-H22		human		M		Lung Tissue		Cauc		Epithelial Cells		ALK-NPM	-100 degree C

Semantic Integration

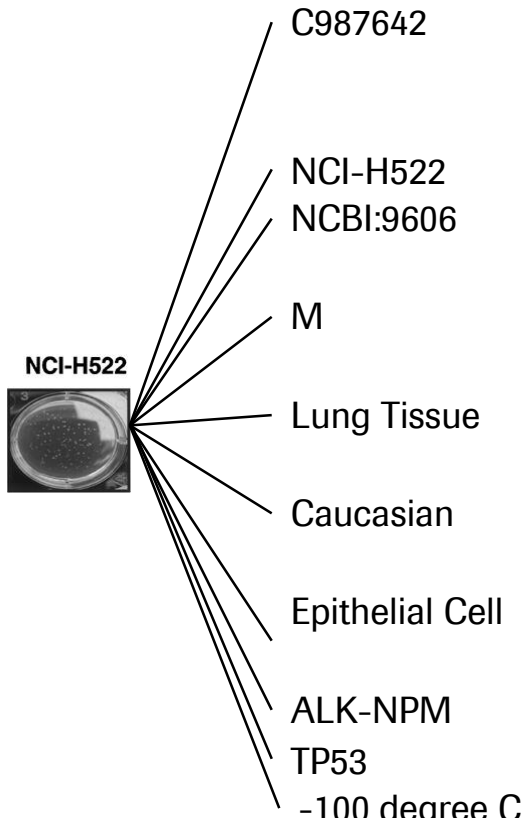
Metadata Alignment (Metadata Registry, Target Schema)



Cell Line ID	Preferred label	Has Species	Sex	Tissue of origin	Ethnicity	Morphology	Mutation	Gene Fusion	Freeze Condition
CVCL1567	NCI-H22	NCBI:9606	Male	Lung	Caucasian	Epithelial Cell	TP53		
CVCL1567	NCI-H22	human	M	Lung Tissue	Cauc	Epithelial Cells		ALK-NPM	-100 degree C

Semantic Integration

Value Domain Integration (Terminology Service)



Cell Line ID	Preferred label	Has Species	Sex	Tissue of origin	Ethnicity	Morphology	Mutation	Gene Fusion	Freeze Condition
CVCL1567	NCI-H22	NCBI:9606	M	Lung	Caucasian	Epithelial Cell	TP53	ALK-NPM	-100 degree C

Semantic Integration

Nice to have or must do?

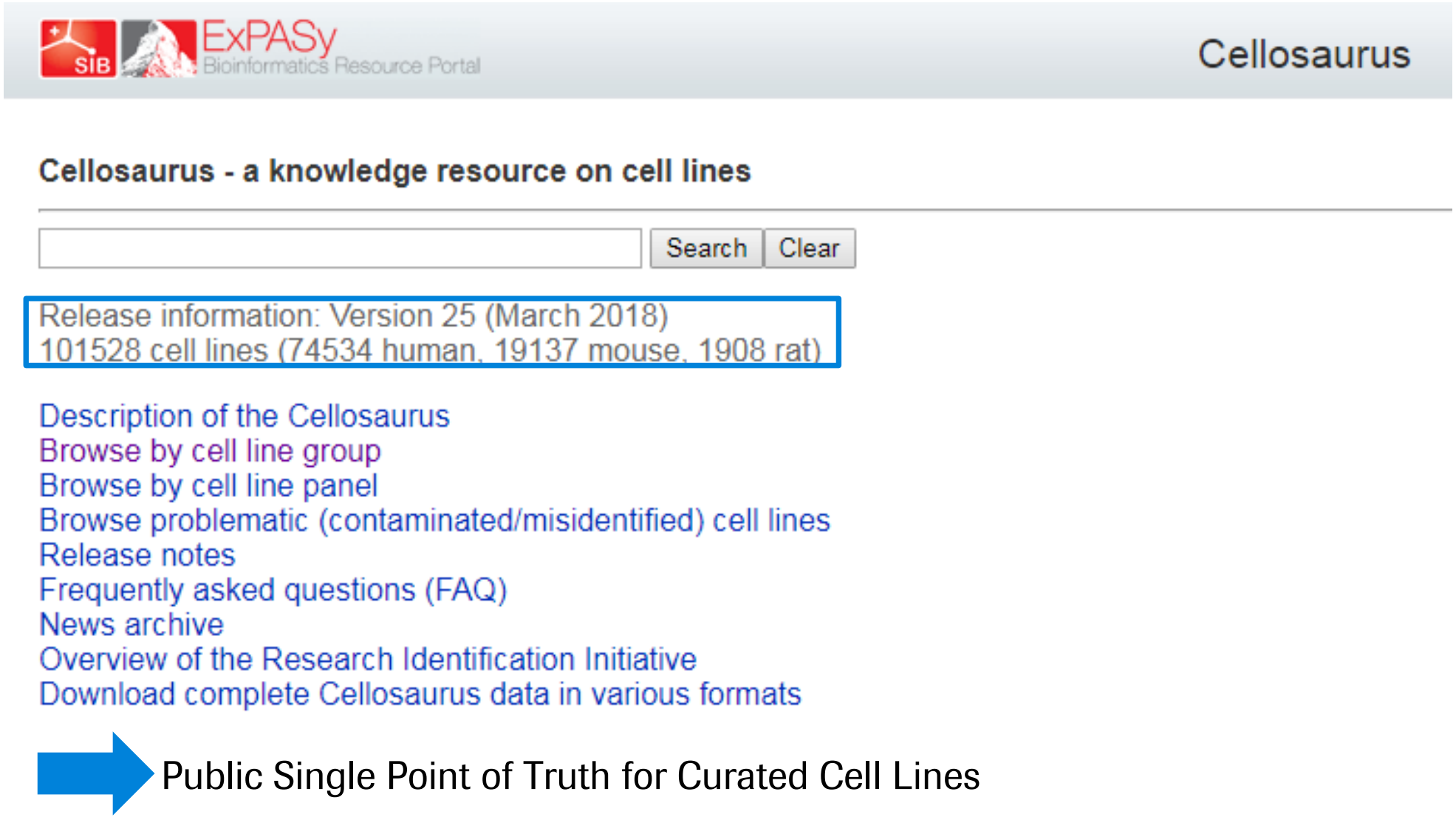
Cell Line ID	Preferred label	Has Species	Sex	Tissue of origin	Ethnicity	Morphology	Mutation	Gene Fusion	Freeze Condition
CVCL1567	NCI-H22	NCBI:9606	M	Lung	Caucasian	Epithelial Cell	TP53	ALK-NPM	-100 degree C

Cell Line ID	ID	Name	Label	Has Species	Species	Gender	Sex	Tissue of origin	Tissue	Ethnicity	Race	Morphology	Cell type	Mutation	Gene Fusion	Freeze Condition
CL234567		NCI-H22		NCBI:9606		Male		Lung		Caucasian		Epithelial Cell		TP53		
	C987642		H522		human		M		Lung Tissue		Cauc		Epithelial Cells		ALK-NPM	-100 degree C

But: Why are *we* doing this?

Public Cell Line Registry

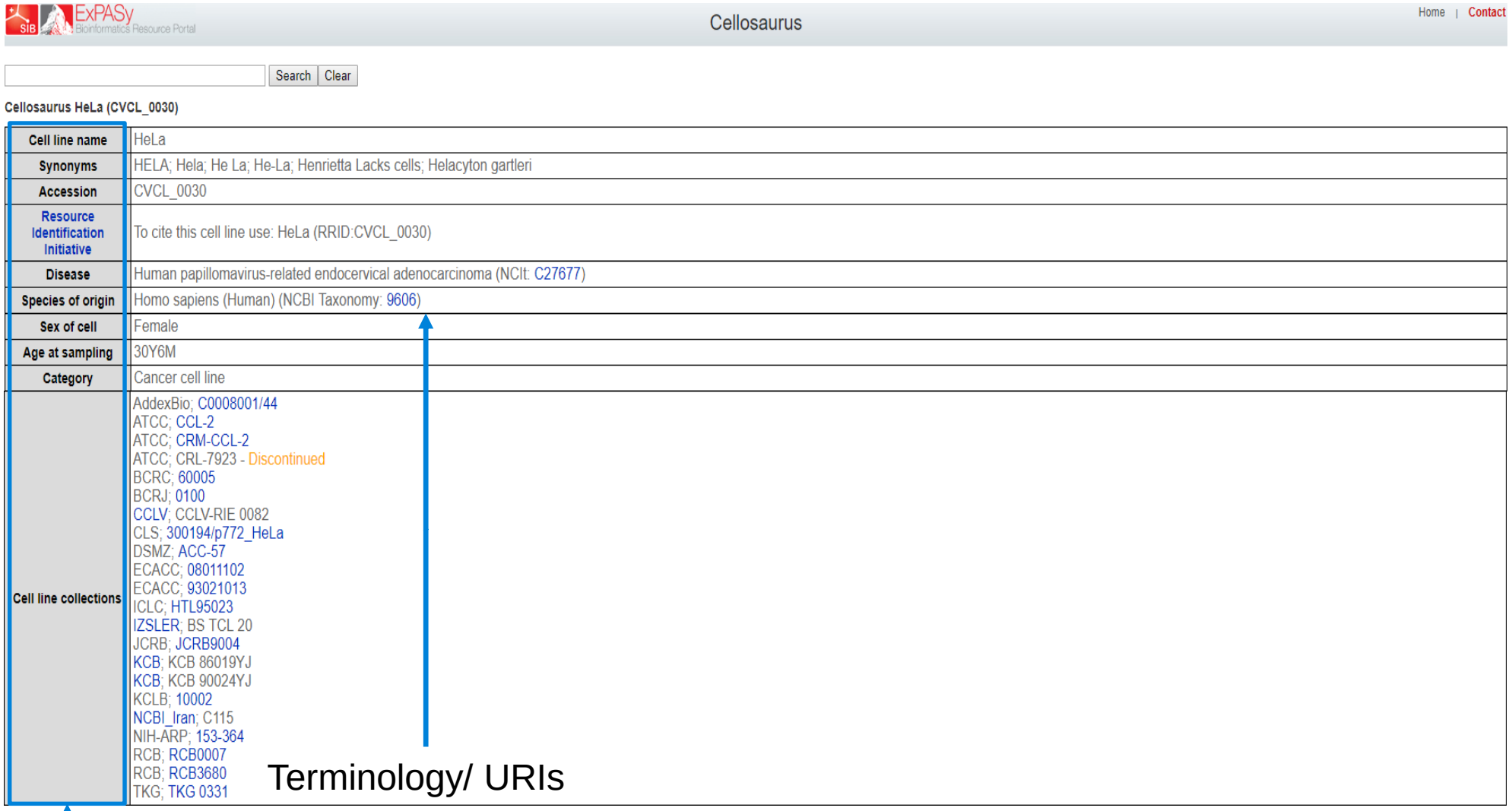
SIB-Cellosaurus (Amos Bairoch)



Public Single Point of Truth for Curated Cell Lines

CelloSaurus

Semantic Organization



Public Cell Line Registry

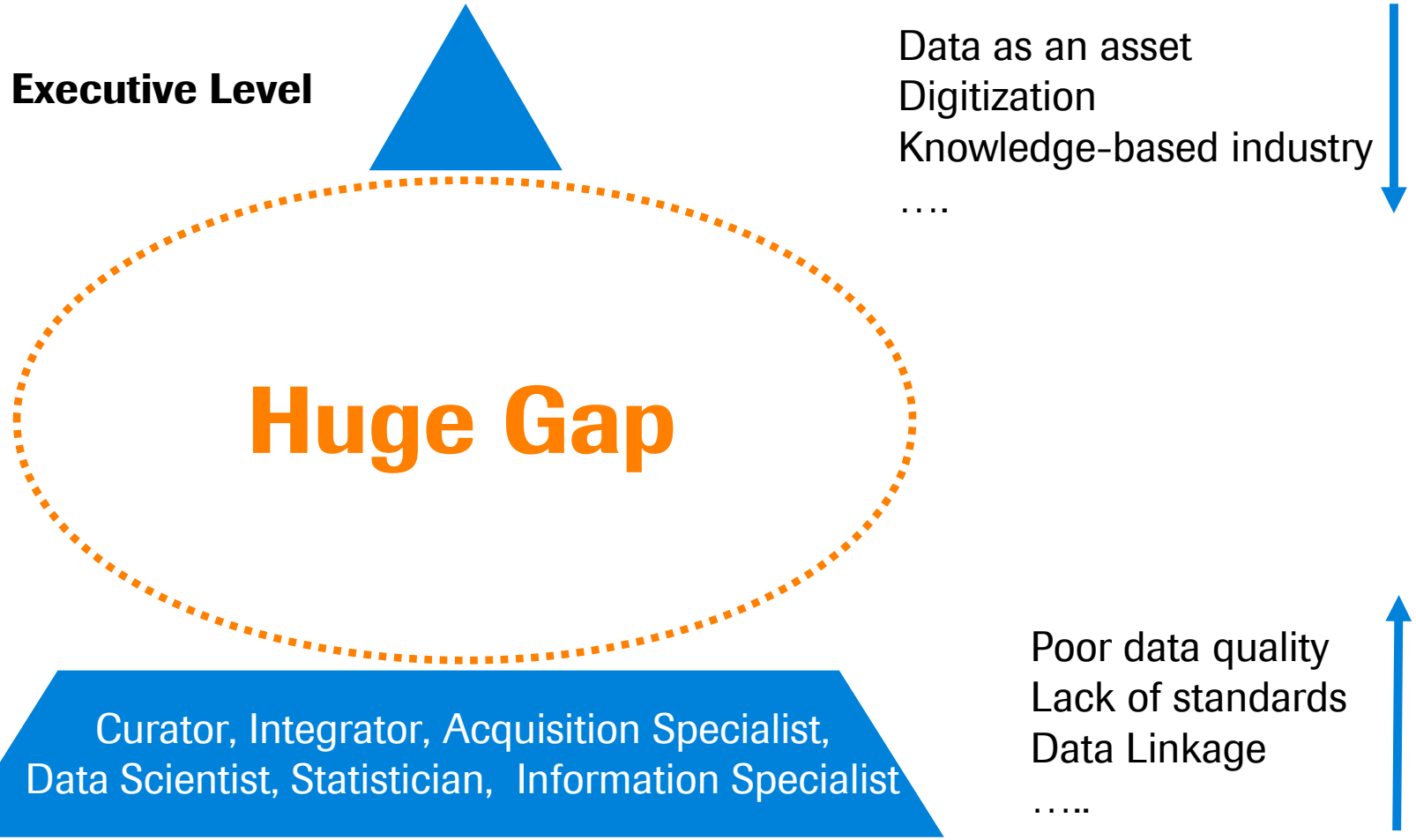
Project Plans & Value Proposition

- Development of a shared Minimal Model for Cell Line Annotation (Metadata)
- FAIRification of Cell Line Registry:
  - Metadata using unique IDs (eg Uberon, NCBI Taxonomy, NCIt)
  - Definition of Data Standards for knowledge represention (Interoperability)
  - Definition of an open protocol for querying/ accessing Cell Line Registry (eg RESTful, JSON, JSON-LD)
- Development of an open collaborative curation platform including vendor data
- Custom partitions for Universities and SMEs to support Cell Line Registration

Examples:
Antibody: RRID:AB_2178887
Cell Line: RRID:CVCL_0033
Organism: RRID:MGI:3840442
Tool: RRID:SCR_007358

FAIRification

Organizational Perspective



Conclusions

- Data-driven industry: Data considered as an asset – support by Executive Managers
- Increasing complexity of data creation, data acquisition and integration processes require Information Procurement based on FAIR principles
- FAIR data is AI ready !
- FAIR principles as guidance for a global data management strategy:
  significant change to the organization (large time horizon)
- Digital Transformation – Hypothesis 1:
  Without Data FAIRification Digital Transformation will create new digital garbage
- Digital Transformation – Hypothesis 2:
  Rapid, cross-funtioncal and global adoption of Data Management based on FAIR principles will create a competitive advantage

Acknowledgements

Acknowledgements

- **pREDi SDA**: Joachim Rupp, Andreas Thalhammer, Fabien Richard, Silvia Jimenez, Felix Schwagereit, David Herzig, Eugen Ulrich, Ludovic Sternberger, Jörg Schmiedle, Marielle van de Pol
- **pREDi Data Science**: Jan Küntzer, Tom Quaiser, Pascal Kuner, Mathias Leddin, Sebastian Scherf, Ralf Jäger
- **pREDi**: Martin Erkens, Michael Braxenthaler
- **gRED**: Stephen Owens, Stephen Day, Christina Lu, Hongmei Huang
- **PD**: Rama Balakrishnan, Petra Strücker, Philipp Ernst, John Franchino, Nelia Lassiera, Ivan Robinson, Kimberly Barnholt, Jonathan Chainey
- **PTD**: Etzard Stolte
- **Pistoia FAIR**: Tom Plasterer, John Wise, Rafael Jimenez, Lars Greiffenberg, Eric Little, Ted Slater, Carmen Nitsche, Drashtti Vassant, Rainer Winnenburg, Alexandra Grebe de Barron, Ian Harrow
- **CelloSaurus**: Amos Bairoch

Doing now what patients need next