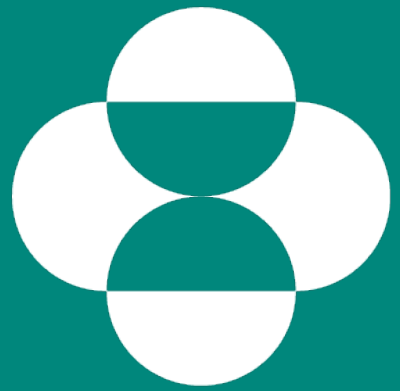
REAL-WORLD CHALLENGES OF LINKED DATA IN LARGE ENTERPRISES

Michael Wurst, PhD
MSD IT, Global Software Engineering, AI



MSD

What’s behind the current AI Boom

Computational power, consumer electronics, network

Some new algorithms and tools

Availability of large amounts of structured, semi-structured and unstructured data

MSD

Can we do the same for the Enterprise?

Why not use these technologies as general purpose technology in large enterprises?

How much revenue did we make in Europe without Germany last year?

What is the temperature in the server room?

How many people are in the building right now?

Who was working on AI related projects in our EU branch last year?

MSD

What needs to be in place to make it work in an Enterprise

- Data must exist
- Data must be structured or possible to structure to some extend
- Data must be reliable
- Data description must be available
- Datasets must be linked or linkable
- Data must be up-to-date
- Data must be accessible

Data must exist in the first place

How to prove to business that collection of data makes sense
(without first having it…)

Who is going to collect or create the data? Can it be retro-collected?

Collection of data can be subject to regulation

Create a compelling business case and take data creation seriously

Data must be reliable

Open Data contain errors that are hard to control

Watson’s 85% target to get it right might not be enough…

For business use this might be a serious problem

Government agencies might require very strong QA (e.g. FDA)

Internal QA / Mirror repositories / Contribute to maintenance of open data

MSD

Data must be (at least partially) structured

Huge amounts of data in enterprises is available in form of documents, emails and scans and other unstructured ways

There usually needs to be at least some extraction of this data to make it useful

MSD

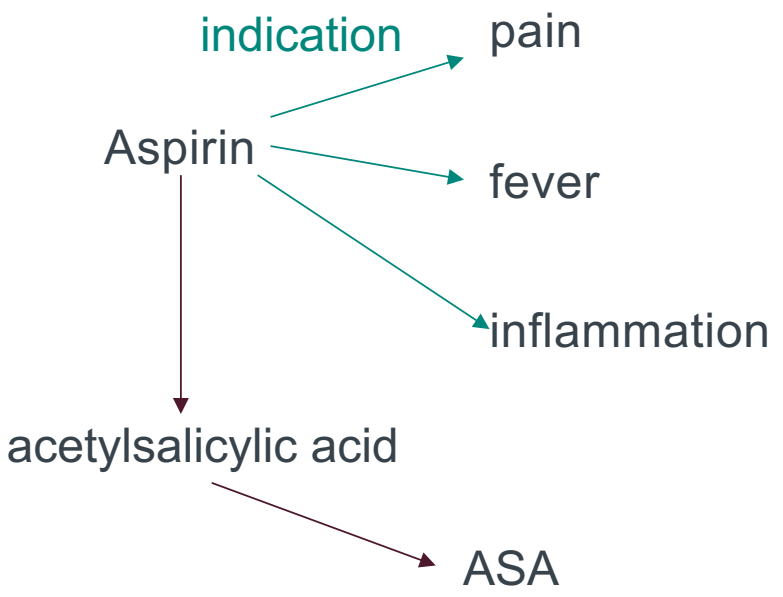
Information Extraction

Extract structured information, like relations, from unstructured or semi-structured sources

Usually based on Natural Language Processing and Machine Learning

Seems like an ideal solution…

Aspirin, also known as **acetylsalicylic acid** (**ASA**), is a medication used to treat pain, fever, or inflammation



Information Extraction

However: Accuracy of models is modest

Most accurate models are based on Machine Learning that needs large amount of labeled samples

This can quickly become rather expensive and requires continuous monitoring/maintenance of models

Evaluate whether information extraction and NLP are applicable at reasonable accuracy/cost point

MSD

(Meta-) Data Without Description

Sometimes there is linked data, however no description or understanding of the semantics

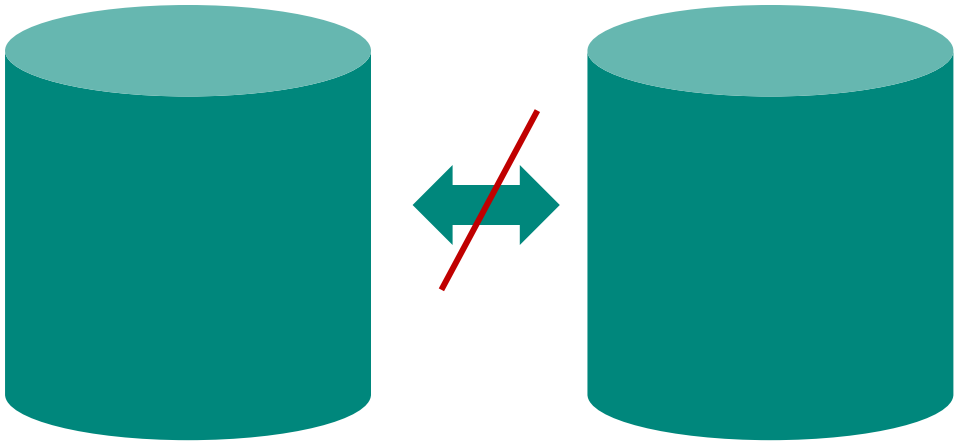
… and nobody wants to or can provide it

Needs close collaboration with business experts

Self-service data science as a potential solution, so that business users can perform part of data processing themselves

MSD

Datasets Must be Linked



Sometimes it is easy, most often it is not

Automatic Schema Mapping, Flexible Databases and Visual Tools can help

However, with commercial data, the problem is sometimes rather the license

MSD

Datasets Must be Updated and Maintained

Some data sources are static, but most change over time

Requirements for just how real-time vary

However, this should not be underestimated when planning a project

Updates and Maintenance should be planned early on

Security and Access

A lot of relevant data is not available for free

Other data, e.g. patient data, is highly sensitive and subject to many regulations

Data owners often avoid risk as they are afraid of possible consequences of data sharing even internally

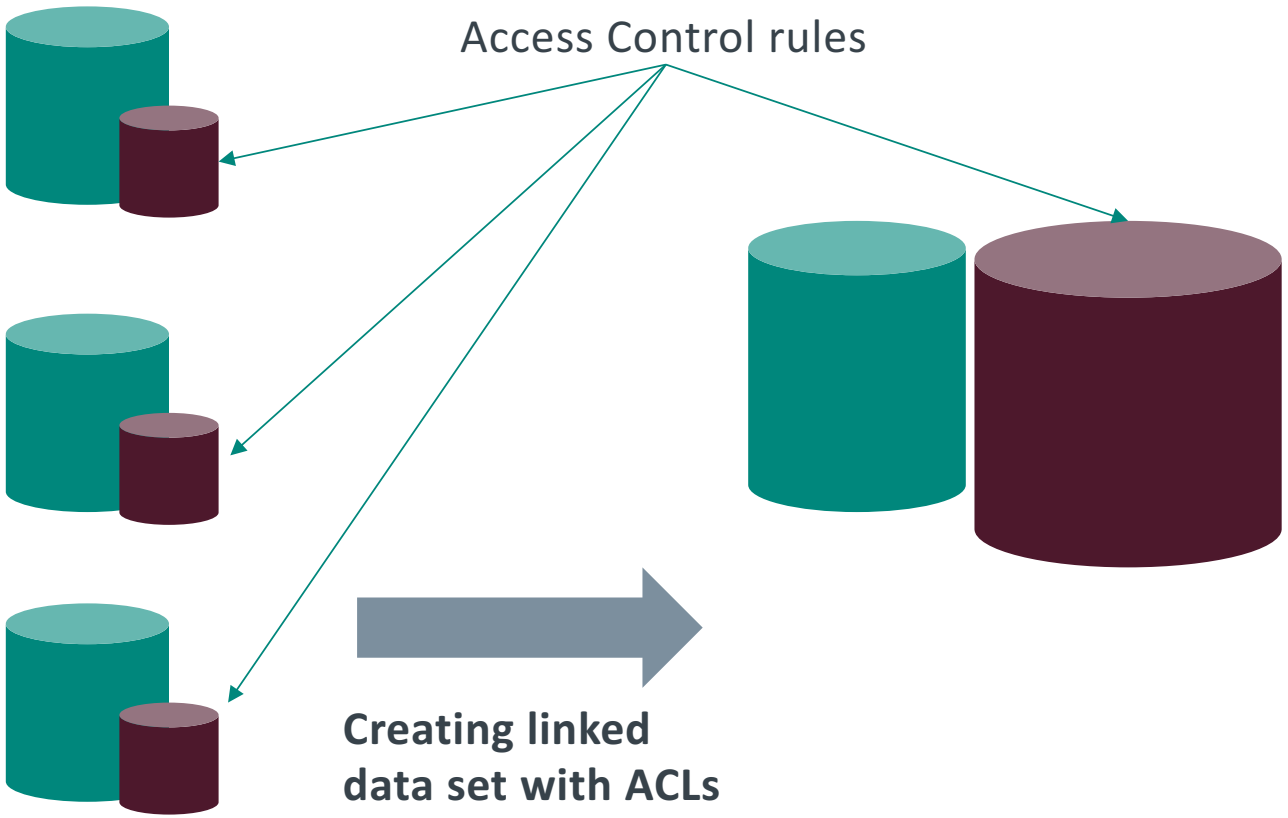
Security and Access

mass	180.042±0 atomic mass unit 1 reference	edit + add value
has effect	acetylsalicylic acid exposure 1 reference	edit + add value
canonical SMILES	CC(=O)OC1=CC=CC=C1C(=O)O 1 reference	edit + add value

This could be highly sensitive!

WikaData: https://www.wikidata.org/wiki/Q18216 (Aspirin)

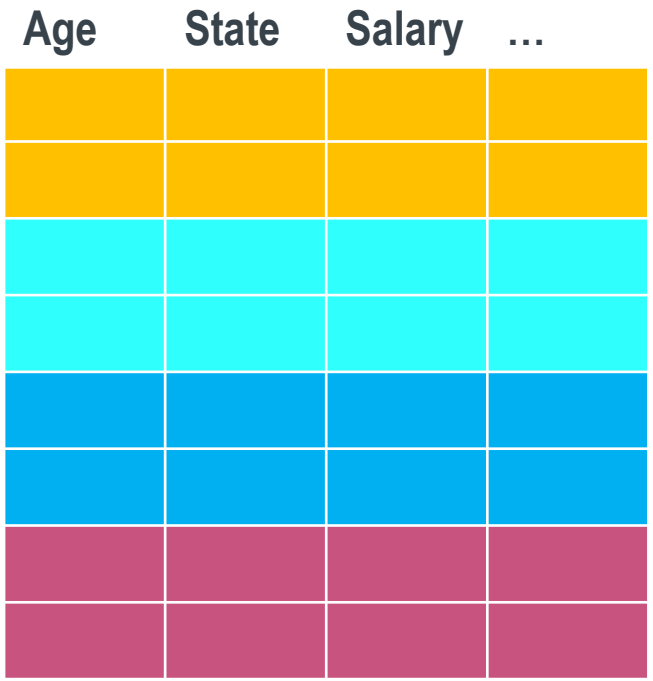
Security and Access



Adhering to all rules on how to combine data and mapping original Access Control to consolidated data can be hard

Application logic gets more complex with clients having access to diverse subsets of the input data

Security and Access – Data Anonymization



One way to get around security and access problems is data anonymization, usually done by removing columns, aggregating rows and/or adding noise

For linked data and meta data, this usually does not work (well)

Why is Enterprise Data Harder than Open Data?

Security/Data Ownership/Complex Access Rights/Licensing/Internal Regulations

Stronger requirements on reliability and control over data and how it is created and updated

Possibility for crowd-sourcing are very limited

Enterprise Data is more strongly tied to legacy applications

Many enterprise applications are very specific and there is no business case to collect and maintain data for it

MSD

Lessons Learned

Be aware of any technology that can help you

Be also aware of the limitations and what they mean for your project

Treat data and meta-data as high risk items in your project
Prototype early, Fail fast

Start small, show value quickly

Plan for data maintenance and updates

Create a culture of sharing data with clear rules and minimal risk for data owners

MSD

Conclusion

(Linked) Data is one of the main limiting factors to implement AI and other technologies in big enterprises

This is only partially a technical problem…

often enough it is not a cultural problem either…

…rather a lack of clear business and compliance rules around data sharing and governance

Key is to identify realistic, sustainable applications based on data and to create a strong business case around them

MSD

Want to join us in solving these and many other problems?

https://www.msdit.cz/studentsgrads

MSD