DEIS Project Vision

The overarching problem identified during the preparation of the BLIZAAR proposal and during the first user workshop (16.-17.2.2016) between LIST and DEIS researchers is that DEIS researchers struggle to get a sense of what information is available in the backend already. The consortium has identified three strategies to address this problem. Each strategy is presented with high-level user actions as well as an outline of a suitable network model. Finally, overarching requirements are listed which need to be considered for all three strategies.

Strategies

1. Diversity and continuous coverage

DEIS researchers need to be able to determine to which extent corpora adhere to internal quality standards. For each entity (person, institution) researchers strive to strike a balance between different types of archives, different types of media (photo, text, video,...) and different political viewpoints. Important entities need to be covered for the full time period of the respective subsection of the corpus. In order to evaluate corpora, researchers need to be able to 1) quickly assess whether these conditions have been met and 2) be made aware of the sections within the corpora where they have not.

For example, there might be a gap in the continuous coverage of the activities of an entity, such as the European Central Bank for 2011. Such gaps are until now usually found by chance or following a specific search for them. Continuous coverage over time is entity-dependent: The European Central Bank was established in 1998, therefore a lack of references to it before this year does not constitute a gap.

High-level tasks

1. See the time period an ePublication and ePublication segments cover (creation/publication date or references to dates in a document)
2. See how many and which archives have been used in an ePublication and in ePublication segments
3. See which document was taken from which archive in an ePublication and in ePublication segments
4. See how many and which media have been used in an ePublication and in ePublication segments
5. See the presence of a given entity across time periods
6. Show the presence of a given entity in ePublications, ePublication segments, documents and co-occurring entities.

2. Search by tag

Researchers would like to have several tags associated with each document and to be able to use

combinations of them to search the corpus. Tags depend on the interests of a researcher, there can therefore not be a definitive or objective catalogue. In contrast to a faceted search, a tag-based approach should allow 1) the discovery of related entities and documents which would be missed by a known-item search and 2) provide a higher-level perspective on a corpus (and the history it represents) as opposed to a document/entity-focus. For example, a search for “Glasnost” should also consider relevant documents about “Mikhail Gorbachev” in the 1980s.

High-level tasks

1. Query for an abstract pattern (consisting of combinations of entities, time range, places) which represents a concrete research interest.
2. Support for the creation of these queries through relevant suggestions (e.g. “Glasnost” → “Gorbatchev”)
3. Explore entities, documents and ePublications which match this pattern
4. Explore entities, documents and ePublications which are related to this pattern

3. Exploration

To get a bird’s eye perspective on the DEIS corpus, an interactive dynamic map of all entities, documents and their relations is needed. There is no one way to grasp and represent the complexity of these relations but several use cases are conceivable: identify(un)wanted gaps between them, understand where a new document, that is about to be added to the system, fits in this map and which gaps it closes, receive recommendations for other documents.

High-level tasks

An interactive dynamic map of all entities, documents and their relations provided by the DEIS. In contrast to the previous strategy, this starts not with a specific query but with a representation of available data. Users identify what they are interested from the general to the specific. The following user actions are conceivable:

1. Explore the overlap and gaps between user-selected searches by tag
2. Explore the overlap, similarities and gaps between DEIS-created themes (ePublications)
3. Identify gaps in continuous coverage
4. Understand where a new document which is about to be added to the Backend fits in this map and which gaps it closes
5. Receive recommendations for other documents
6. Categorize types of recommended documents which allow users to explore other themes associated with their original theme.

Overarching requirements

Context

(Surely not only) in historical analyses is the context in which an observed object is embedded crucial

for its interpretation. In the case of the data we have available, context means entity co-occurrences, document metadata, associated ePublications, time ranges, geography.

Scale dependent zoom

Many of the third-party visualisations I have encountered in the context of BLIZAAR (one example: https://multinets.io/share/politics_ch_twitter/#) show relations between datasets or network layers. Something I regularly found hard in the available demos was to drill down to specific nodes or edges (or subsets) to evaluate them. It is relatively easy to observe large-scale structures and to identify something potentially interesting. It is hard to zoom in to nodes, edges or subsets thereof to verify that an observation is indeed interesting. One could describe this with reference to digital geographical maps: On a large scale structural information of the graph would be displayed which enables a user to identify what is interesting. As they zoom in on a region (a.k.a scale dependent zoom), additional information becomes available which matches well the respective scale.

Document-level exploration

Any observation and pattern needs to be verified through the evaluation of the underlying documents. Any visualisation therefore needs to also support the display of selected documents.

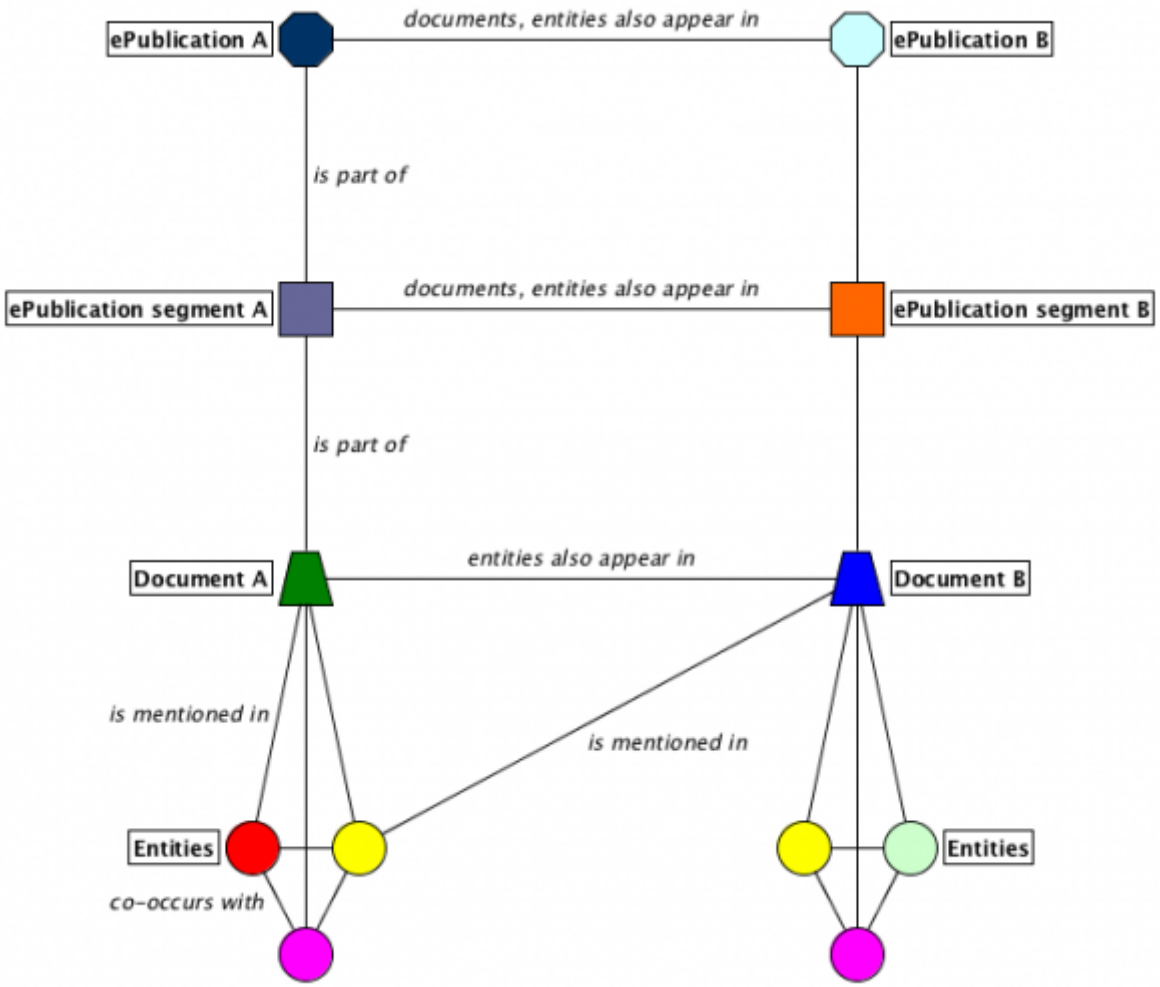
Keep track of actions

Especially in complex visualisations, the path of actions to navigate needs to be recorded. For the following reasons: Reroute after hitting an exploratory dead-end. Ability to comprehend and evaluate the meaning for a created view on the data.

Preliminary multilayer network model

This could be modeled as a multilayer, dynamic network:

- Level 1 shows co-occurrences between the selected entity and other entities (one could distinguish between places, persons, institutions).
- Level 2 shows documents which reference the selected entity and relations between them based on co-occurring other entities (one could distinguish between places, persons, institutions).
- Level 3 shows ePublication segments which reference the entity and relations between them based on co-occurring other documents.
- Level 4 shows ePublications which reference the entity and relations between them based on co-occurring other documents.



LIST Life Sciences Project Vision

The aim of the project is to generate an integrated image of several sets of biological data collected in experimentally related samples (i.e. cannabis samples) in different “omics” experiments: transcriptomics, proteomics and metabolomics. This integrated image will facilitate the building of knowledge in biological studies. The intertalk between genes, which code for proteins, which in turn can regulate genes (and hence themselves) and catalyze the transformation of metabolites is a very complex and finely tuned phenomenon. The network metaphor is effective as a model and might allow biologists to understand the meaning of the collected data, despite the noise of biological processes not directly targeted by the experimental factors and technical limitations (e.g. missing data). Several strategies can be adopted to tackle the problem.

Strategies

1. Pathways reconstruction

Missing data requires filling the gaps between regulated entities, to try to complete the puzzle of the biological process and allow hypothesis making. Proteins act transmitting and transforming the signals (i.e. “transducing” the stimuli) to trigger gene regulation and metabolism. When a link in the

transduction path is missing, filling it might allow not only to trace it, but also to understand collateral regulated paths, as the regulatory networks are scale free in biology. This approach shall be regarded as tentative, because of the complexity of the biological networks, and uncertainty of the results shall always be available and clear to the user. High-level tasks:

1. Find all the first neighbors of two given proteins (i.e. “missing links”)
2. Find enzymes active on a metabolite
3. Find missing links given an experimentally measured threshold
4. Find coherent values across different layers (i.e. experimentally up across different layers)
5. Build a layer using expression data (available through STRING for general experiments or from plant group one), where edges are inferred by similarity in the profiles (distance to be decided)
6. Given a threshold, suggest to the user how richer or poorer the network would become by adjusting it (e.g. showing a distribution of edges amount in the graph in function of the threshold)

2. Pathway analysis

Once the regulatory network has been sketched, it shall be analyzed; this analysis results in the identification of key molecules, genes, proteins and metabolites. This analysis can also serve as experimental data confirmatory step, where up and down regulations follow identified paths. Because of the aforementioned collateral pathways, the complexity of the network shall be reduced, by selecting significant interactions. High-level tasks:

1. Identify key nodes shortest paths (nature *tends* to favor the quickest transduction way available)
2. Find enzymes with data form both the proteomics and transcriptomics and also for the substrates (metabolites)
3. Sort key molecules, via centrality measures between given ends
4. Track the path from a receptor to transcription factors and vice versa
5. Use Gene Ontology to subset the network and keep only significant proteins and genes

3. Exploration

Global visualization of regulated genes, proteins and metabolites, despite its complexity, it offers a valuable tool for hypothesis generation and insights elicitation. Mapping experimental values and ontology classification to the network and trigger reorganization base on them would allow a better understanding of the biological phenomenon. High-level tasks:

1. Show and lay out nodes based on layer and highlight cross-layer connections
2. Identify up- and down-regulated regions based on homogeneity of experimental values
3. Identify a metabolic pathway, show underlying regulations
4. Understand the consequences of a hypothetical change on other layers (e.g. remove a node, to simulate a knockout)
5. Implement in a DOI function a measure able to grasp and identify molecules behaviors such as moonlight proteins (i.e. proteins with different functions and interacting molecules, in different conditions)
6. Use GO to explore proteins location, in particular proximity and how this correlates with experimental data
7. Clustering by expression profiles

8. Integrate other analytical approaches such as parallel coordinates

Overarching requirements

Scale dependent zoom

If I understood correctly the humanities case, this would imply some level of grouping while zooming out. Gene ontology might be used to group transcriptomic and proteomic nodes into groups, while intersections might render this challenging, but this remains an attractive idea.

Layout

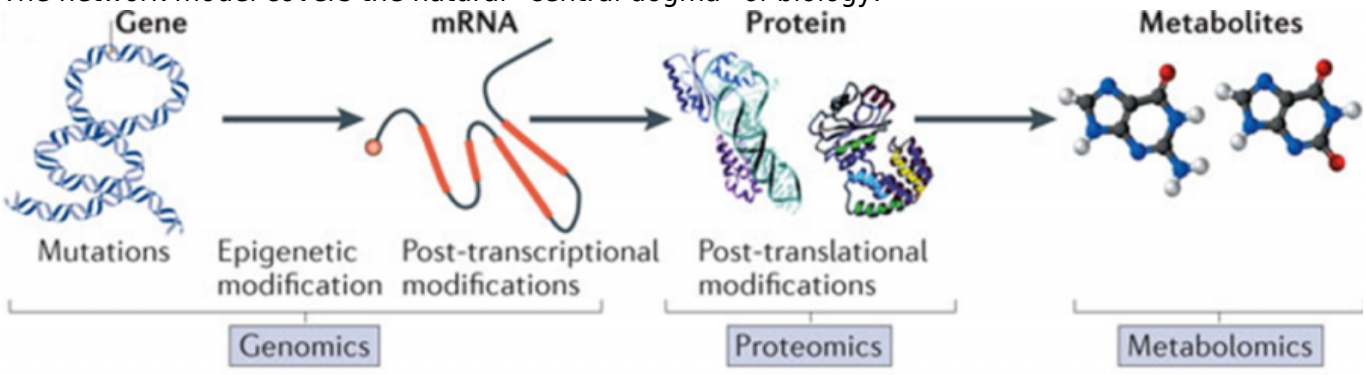
A layout base on experimental values, on ontology or network features such as centrality, and at the same time aware of the other layers, might be interesting for the exploration.

Keep track of actions

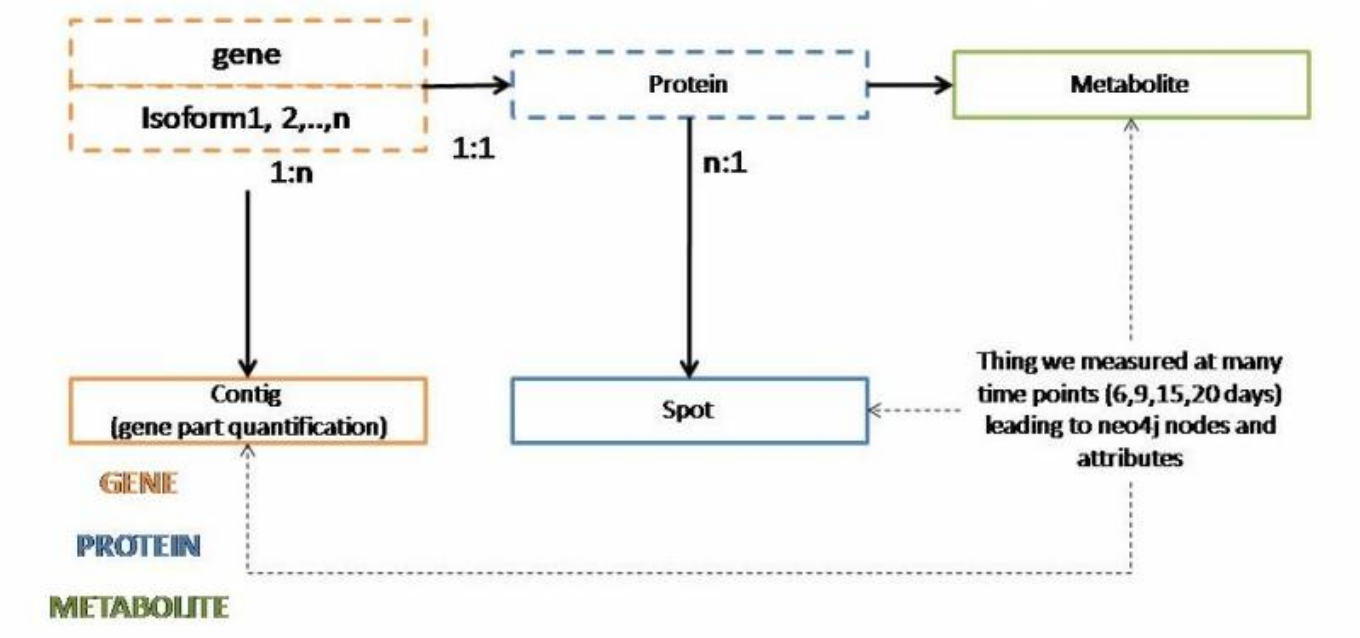
As in the humanities case, the possibility to keep track and reproduce steps might be very useful.

Preliminary multilayer network model

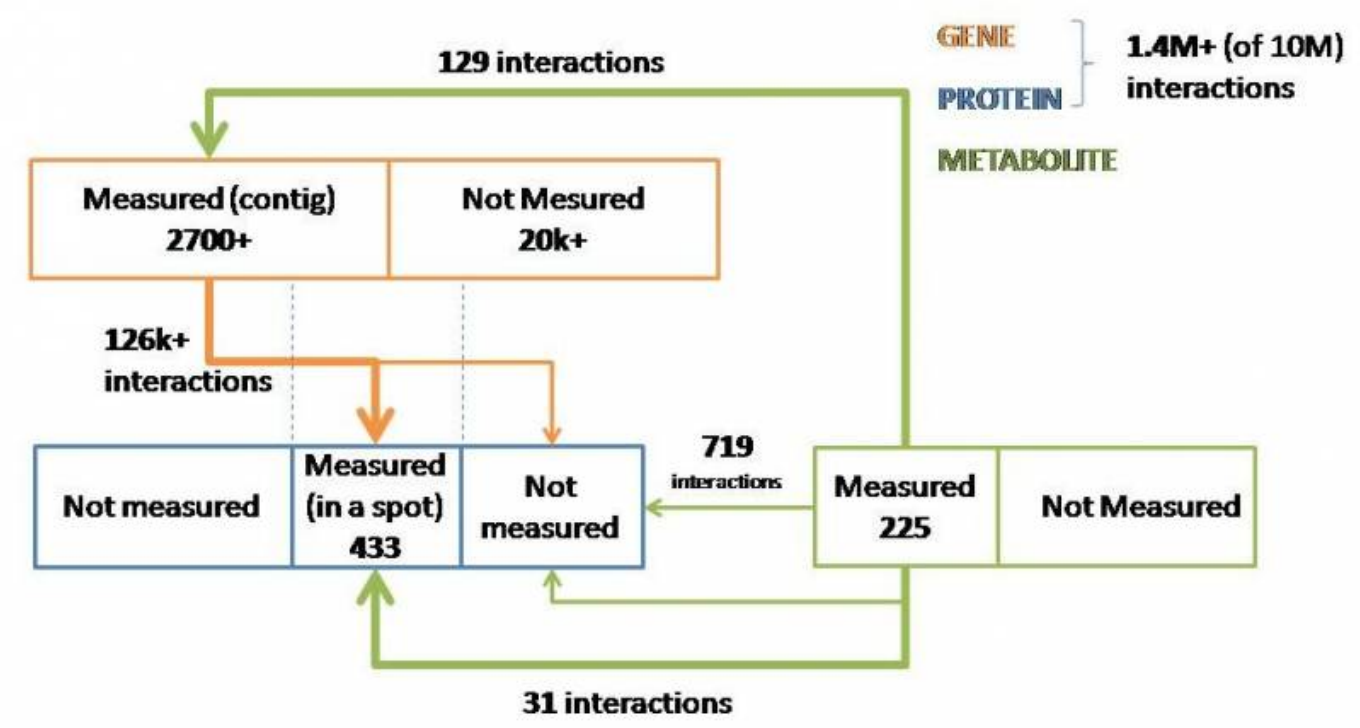
The network model covers the natural “central dogma” of biology:



With each layer corresponding to a “omic” domain.



Entities and measure on them



Relationships among measured entities (bold arrows) and non measured ones, yet in the database (thinner lines)

From:
http://blizaar.list.lu/ - **BLIZAAR**

Permanent link:
http://blizaar.list.lu/doku.php?id=project_vision

Last update: **2016/11/11 14:26**