Extracting value from grey literature

Processes and technologies for aggregating and analysing the hidden Big Data treasure of the organisations

Big data: the 3 "V"s

Data sets satisfying one or more of the following requisites:

- **Volume**: huge amounts of data, which cannot be effectively managed with usual technologies (e.g.: relational data base management systems)
- **Velocity**: high throughput data collection and provisioning
- **Variety**: heterogeneity of sources, types and formats

Big data: definition issues 1/2

- Need for a **less ambiguous operational definition**, with explicit reference to usage contexts, technological environments and involved actors
  - e.g.: Volume and Velocity thresholds cannot be defined as absolute values because they are strictly connected to current technological constraints

- Current definition does not take into account **all challenges**
  - What about
    - **Veracity**: how can I ensure data reliability in such an heterogeneous and dynamic scenario?
    - **Validity**: how can I ascertain relevance for the intended use?
    - **Volatility**: how long is data valid and how long should it be stored?

Big data: definition issues 2/2

- **Volume, velocity, variety** → quantitative and measurable aspects, more easily definable
- **Veracity, volatility, validity** → cannot be assessed with a simple, direct measure.

Big data and grey literature

In grey literature, big data challenges include the management and processing of:

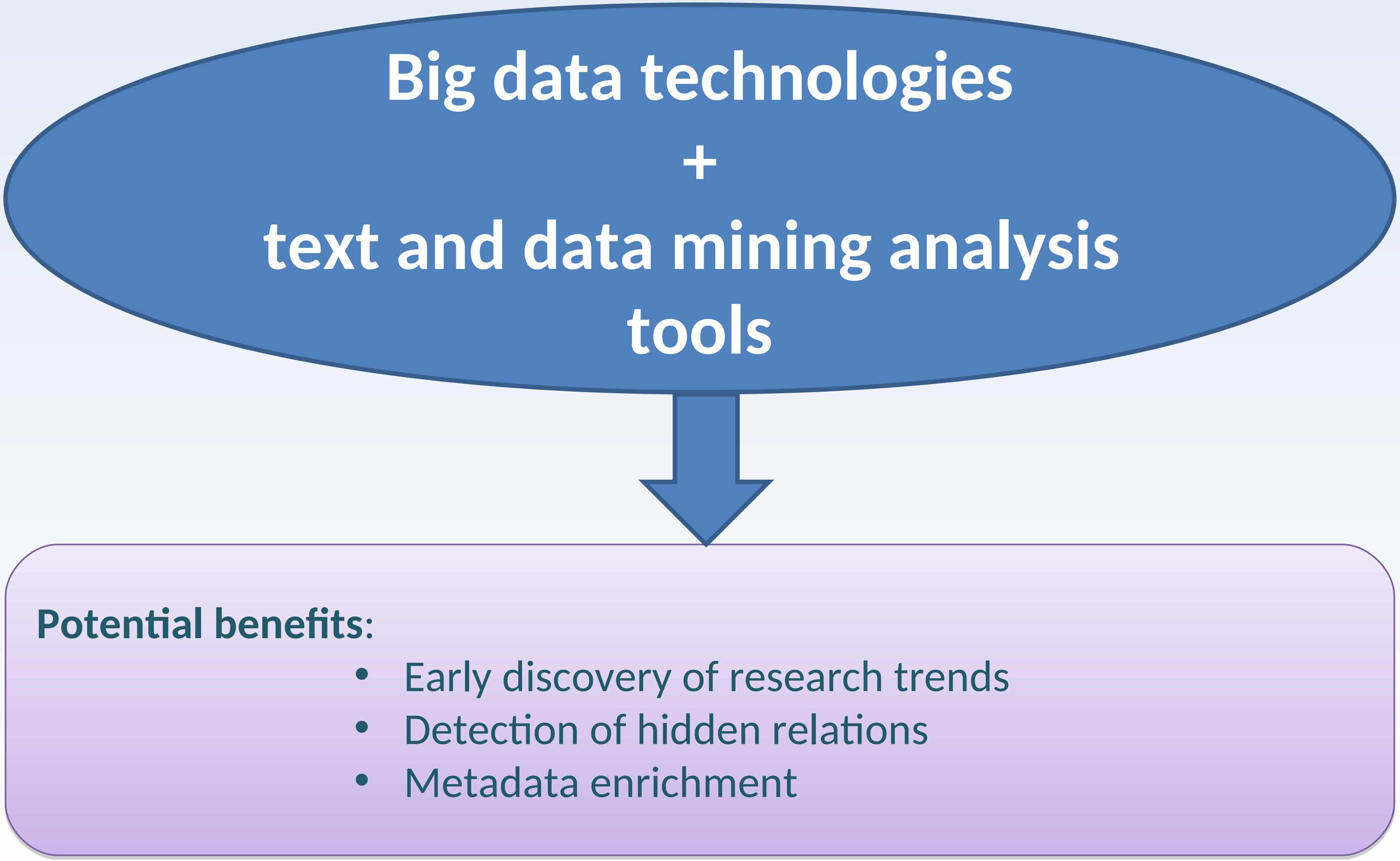
- **Digital contents**
- **Metadata**
- **Contexts and relations**

Grey literature products:

•are by definition characterized by **Variety** – in terms of heterogeneity of content types, formats and internal structures

•present issues of **Veracity**, **Volatility** and **Validity**

Grey literature: text and data mining on a large scale



Big data technologies
+
text and data mining analysis tools

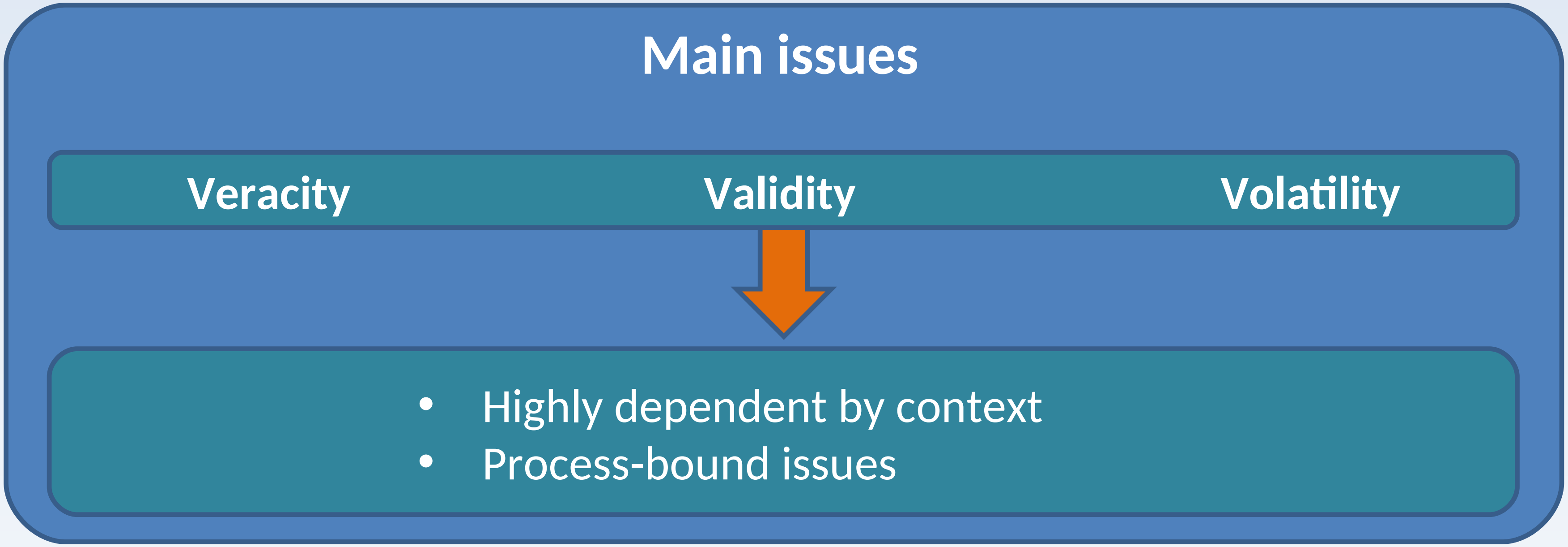
Potential benefits:

- Early discovery of research trends
- Detection of hidden relations
- Metadata enrichment

Stakeholders

- **Political decision-makers**: support for long term research planning
- **Industry**: useful indicators to be taken into account for future investments
- **Researchers**: detection of upcoming/implicit research trends or interesting connections between research groups/fields

Issues and solutions



Main issues

Veracity Validity Volatility

- Highly dependent by context
- Process-bound issues

Proposed solutions:

- quality level agreements
- data cleansing procedures
- cross-check with external sources.

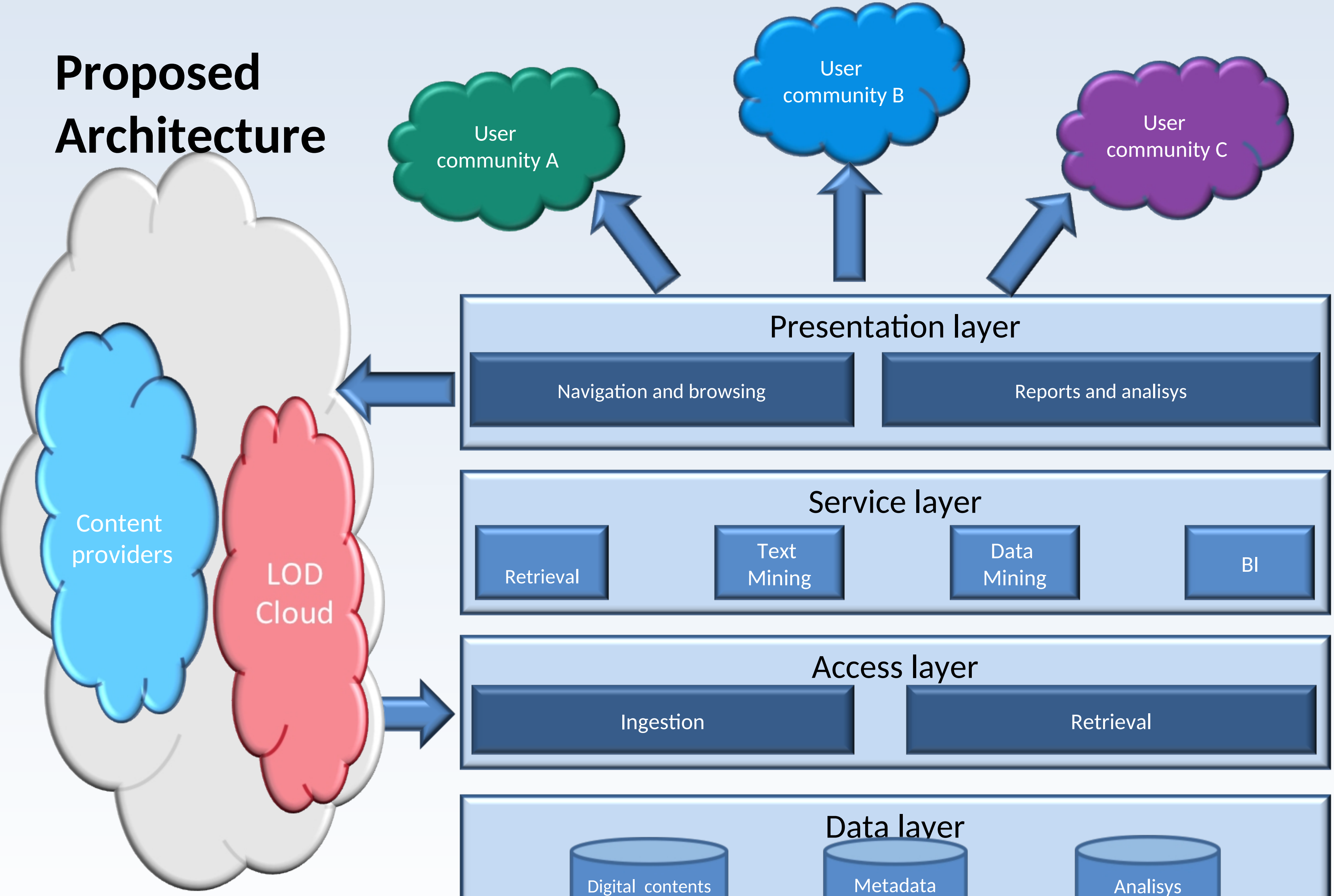
To Be Addressed:

- Validity
- Volatility

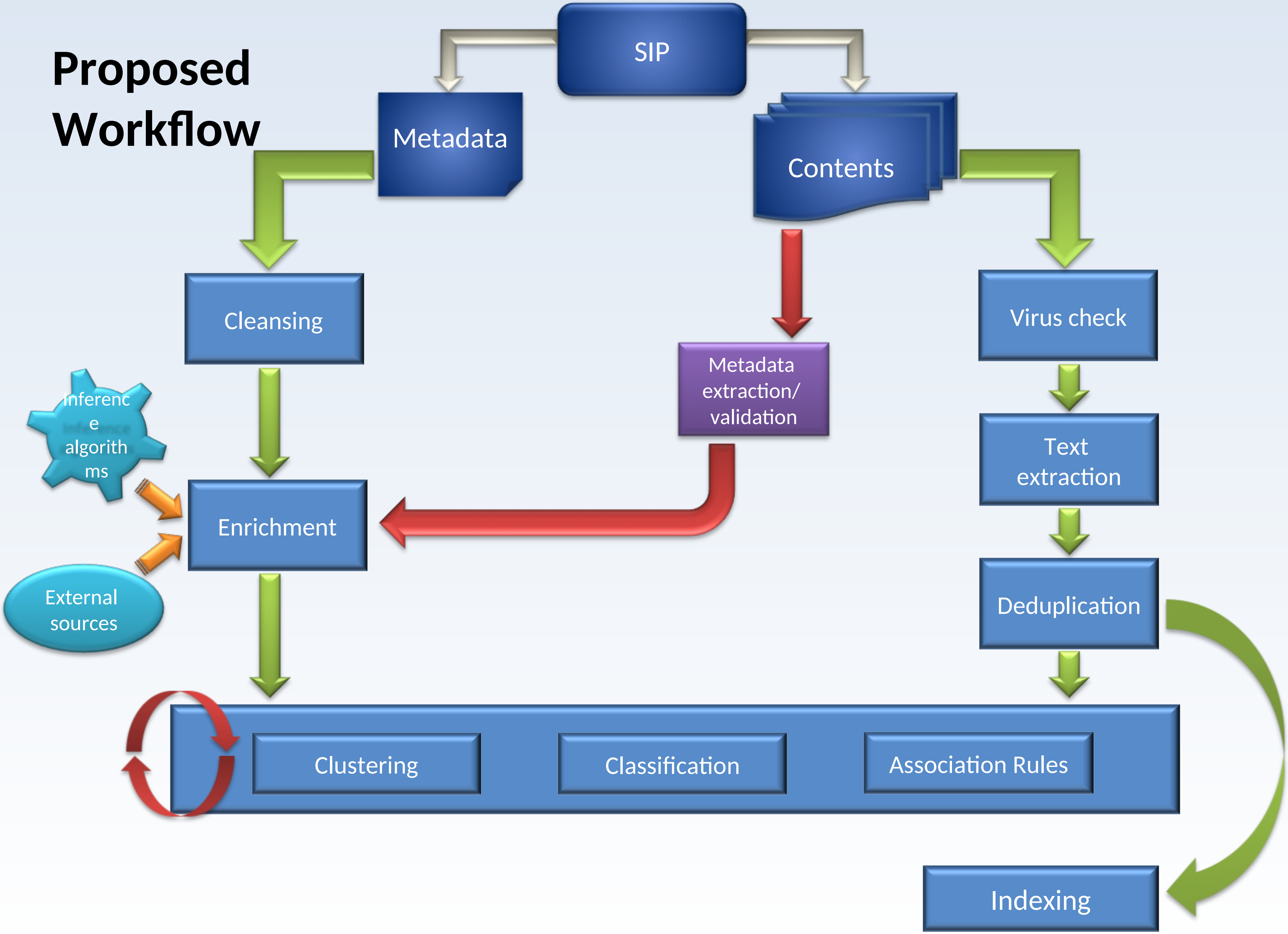
Analysis types

- **Classification** → e.g. for metadata enrichment and linking
- **Clustering** → e.g. detection of non-trivial connections between research fields or organizations/research groups
- **Association rules** → pattern detection (e.g. if a research group is specialized in field A, they have or will develop connections with groups specialized in field B with probability *x)*

Proposed Architecture



Proposed Workflow



SIP

Metadata

Contents

Cleansing

Metadata extraction/ validation

Virus check

Inference algorithms

Enrichment

Text extraction

External sources

Deduplication

Clustering

Classification

Association Rules

Indexing

Conclusions and future work

- Main issues mostly related to **organisational aspects**
- Benefits from the integration of metadata, contents and analysis results in the **LOD cloud** → system's value added
- Importance of **feedback from user communities** for the continuous improvement of analysis result quality
- Formal definitions for **Veracity**, **Validity**, **Volatility** and related evaluation criteria.