

2.4.2 Information Extraction and Observation

[Return to 2.4 Existing Functional Capabilities in Reporting Systems](#)

Reporting systems typically include a functional capability that derives machine-processable information from submitted source artefacts. This capability encompasses extraction techniques such as optical character recognition (OCR), structured parsing, natural language processing (NLP), and artificial intelligence-assisted methods. These techniques identify values, structures, and other features in reported materials [\[R8\]](#), [\[R9\]](#).

The primary responsibility of this functional component is to produce observed facts: information that a system detects or extracts from reported artefacts. Observed facts remain distinct from claims made by a reporting party and from subsequent interpretations produced through [semantic](#) rules, calculations, classifications, or contextual logic [\[R11\]](#).

Where extraction operates on unstructured or semi-structured sources, observed facts may include:

- Uncertainty measures
- Confidence scores
- Extraction methods
- Source locations
- Processing timestamps
- Model or tool identifiers
- Other extraction context

Although extraction technologies can produce machine-processable outputs, implementations often embed the meaning of extracted information implicitly within code, models, prompts, configuration, or tool-specific conventions. This coupling entangles observation with interpretation and makes it difficult to:

- Trace how a system derived an observed value
- Distinguish extraction behaviour from interpretive logic
- Reassess an observation under revised semantic definitions
- Compare results produced by different extraction techniques
- Isolate extraction errors from downstream validation or analysis

These concerns reinforce the need for explicit architectural separation between observation and interpretation [\[R10\]](#), [\[R11\]](#).

From an architectural perspective, this functional capability concerns what a system observes from a source artefact. It does not determine how downstream components interpret, validate, classify, calculate, or assess that information.

The [Reference Architecture](#) therefore treats information extraction and observation as responsibilities distinct from interpretation and validation. This separation supports traceability, auditability, reproducibility, and controlled semantic evolution throughout the reporting lifecycle, even as extraction technologies and artificial intelligence-assisted methods continue to evolve [\[R10\]](#).

From:
<https://wiki.didosolutions.com/> - **Dido Solutions, Inc Wiki**

Permanent link:
<https://wiki.didosolutions.com/dido/01-fdis-ra/02-background-and-motivation/02-4-existing-functional-capabilities-in-reporting-systems/02-4-1-source-artifact-persistence/start?rev=1783881402>

Last update: **2026/07/12 11:36**

