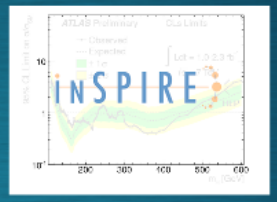
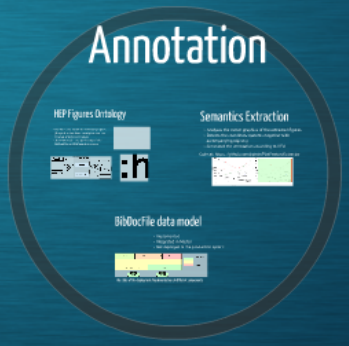
# The Infrastructure for Figures in INSPIRE

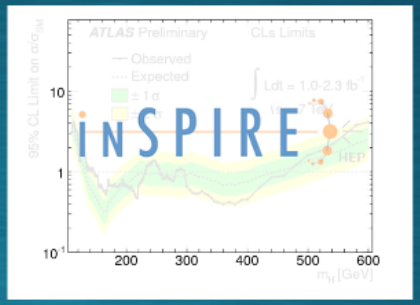


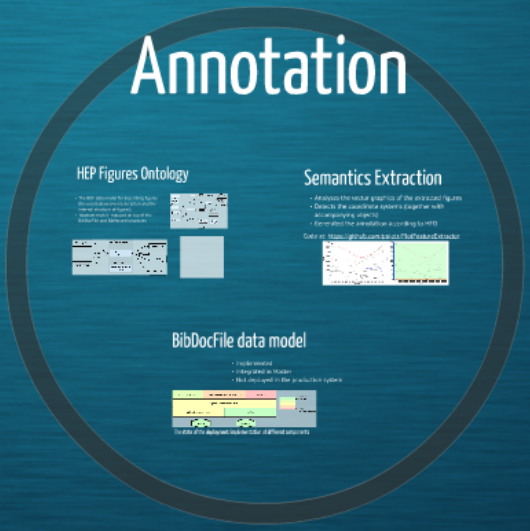




Piotr Praczyk (piotr.praczyk@gmail.com) CERN, 27.08.2013

# The Infrastructure for Figures in INSPIRE

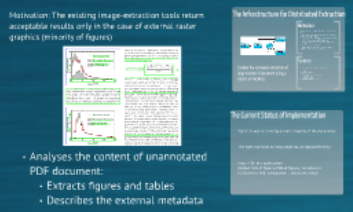





Piotr Praczyk (piotr.praczyk@gmail.com) CERN, 27.08.2013

# Acquisition

## Sources of Documents




- Analyses the content of unannotated PDF document:
  - Extracts figures and tables
  - Describes the external metadata

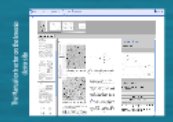
## PDF Extractor

## LaTeX Extractor

## Extractor Merger

- Accepts the output of PDF and LaTeX extractors
- Detects the same figures in both outputs
- Merges the annotation to create a single annotation which is better than separate sources

Code at: https://github.com/poiotr/invenio/tree/docextract
( to be rebased, squashed, tested with the new Invenio)



- Allows to browse the publication page by page
- Allows to select figures and captions
- Attempts to automatically extract text of captions
- Rotations and scaling allow to correctly extract figures from scanned document

## Manual Extractor

## INSPIRE HEP

Motivation: The existing image-extraction tools return acceptable results only in the case of external raster graphics (minority of figures)

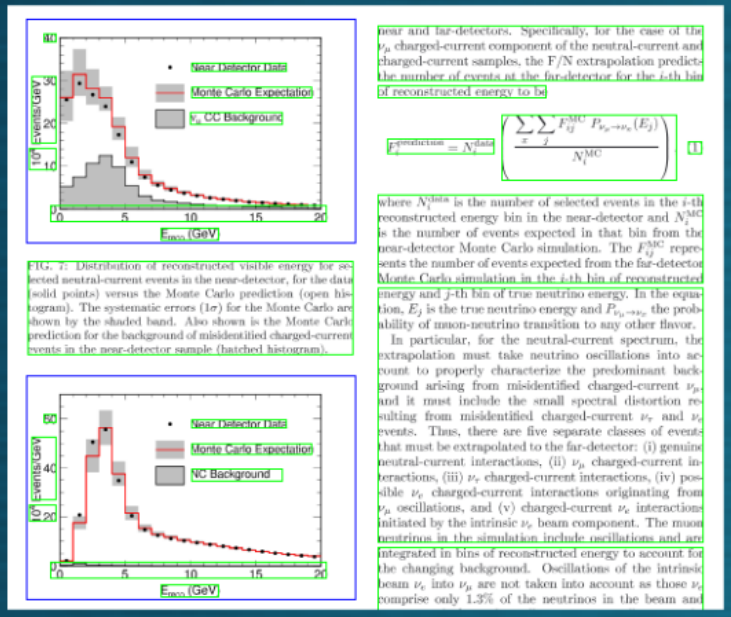



FIG. 7: Distribution of reconstructed visible energy for selected neutral-current events in the near-detector, for the data (solid points) versus the Monte Carlo prediction (open histogram). The systematic errors ($1\sigma$) for the Monte Carlo are shown by the shaded band. Also shown is the Monte Carlo prediction for the background of misidentified charged-current events in the near-detector sample (hatched histogram).

near and far-detectors. Specifically, for the case of the $\nu_\mu$ charged-current component of the neutral-current and charged-current samples, the F/N extrapolation predicts the number of events at the far-detector for the $i$-th bin of reconstructed energy to be

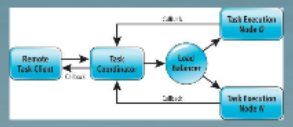
$$F_i^{\text{prediction}} = N_i^{\text{data}} \left( \frac{\sum_x \sum_j F_{ij}^{\text{MC}} P_{\nu_\mu \to \nu_e}(E_j)}{N_i^{\text{MC}}} \right) \quad (1)$$

where $N_i^{\text{data}}$ is the number of selected events in the $i$-th reconstructed energy bin in the near-detector and $N_i^{\text{MC}}$ is the number of events expected in that bin from the near-detector Monte Carlo simulation. The $F_{ij}^{\text{MC}}$ represents the number of events expected from the far-detector Monte Carlo simulation in the $i$-th bin of reconstructed energy and $j$-th bin of true neutrino energy. In the equation, $E_j$ is the true neutrino energy and $P_{\nu_\mu \to \nu_\mu}$ the probability of muon-neutrino transition to any other flavor.

In particular, for the neutral-current spectrum, the extrapolation must take neutrino oscillations into account to properly characterize the predominant background arising from misidentified charged-current $\nu_\mu$, and it must include the small spectral distortion resulting from misidentified charged-current $\nu_\tau$ and $\nu_e$ events. Thus, there are five separate classes of events that must be extrapolated to the far-detector: (i) genuine neutral-current interactions, (ii) $\nu_\mu$ charged-current interactions, (iii) $\nu_\tau$ charged-current interactions, (iv) possible $\nu_e$ charged-current interactions originating from $\nu_\mu$ oscillations, and (v) charged-current $\nu_e$ interactions initiated by the intrinsic $\nu_e$ beam component. The muon neutrinos in the simulation include oscillations and are integrated in bins of reconstructed energy to account for the changing background. Oscillations of the intrinsic beam $\nu_e$ into $\nu_\mu$ are not taken into account as those $\nu_e$ comprise only 1.3% of the neutrinos in the beam and

- Analyses the content of unannotated PDF document:
  - Extracts figures and tables
  - Describes the external metadata

## The Infrastructure for Distributed Extraction




Enables the automatic extraction of large number of document using a cluster of machines

### Motivation

- Extraction of over 1 megapixel requires more than a single machine
- Adding dependencies (like existing scheduling framework) is too painful for such a project
  - It might be difficult to deploy a batch framework in a very restricted CERN environment (bureaucracy)
  - We have super-low priorities in the public batching system
- It was easy and fun to implement
- I needed large-scale extraction during the development (tuning the parameters of the algorithm)

### Features

- Distributed execution on a (dynamic) cluster of machines
- Task-level failover
  - Timeouts/error handling
  - Resubmission
- No installation required

## The Current Status of Implementation

"beta" - it works correctly on vast majority of the documents

The code available at: https://github.com/ppiotr/Invenio

Input: PDF of a publication
Output: SVG of figures, PNG of figures, metadata in JSON and in XML (redundant ... one to be killed)


# PDF Extractor

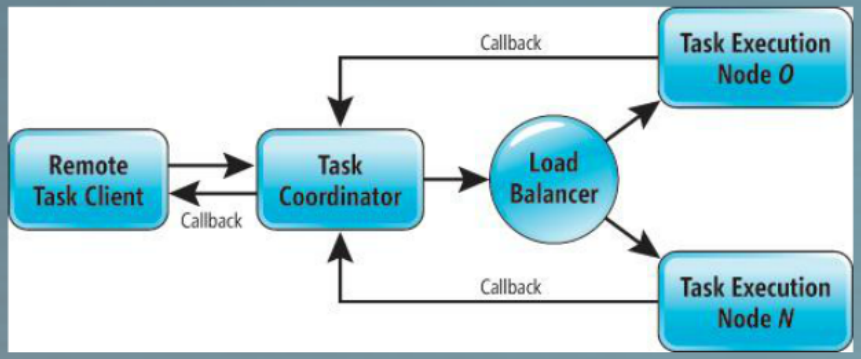
# The Current Status of Implementation

"beta" - it works correctly on vast majority of the documents

The code available at: https://github.com/ppiotr/Invenio

Input: PDF of a publication
Output: SVG of figures, PNG of figures, metadata in JSON and in XML  (redundant ... one to be killed)

# The Infrastructure for Distributed Extraction




## Motivation

- Extraction of over 1 megafigure requires more than a single machine
- Adding dependencies (like existing scheduling framework) is too painful for such a project
  - It might be difficult to deploy a batch framework in a very restricted CERN environment (bureaucracy)
  - We have super-low priorities in the public batching system
- It was easy and fun to implement
- I needed large-scale extraction during the development (tuning the parameters of the algorithm)

## Features

- Distributed execution on a (dynamic) cluster of machines
- Task-level failover
  - Timeouts/error handling
  - Resubmission
- No installation required

Enables the automatic extraction of large number of document using a cluster of machines

# Motivation

- Extraction of over 1 megafigure requires more than a single machine
- Adding dependencies (like existing scheduling framework) is too painful for such a project
  - It might be difficult to deploy a batch framework in a very restricted CERN environment (bureaucracy)
  - We have super-low priorities in the public batching system
- It was easy and fun to implement
- I needed large-scale extraction during the development (tuning the parameters of the algorithm)

# Features

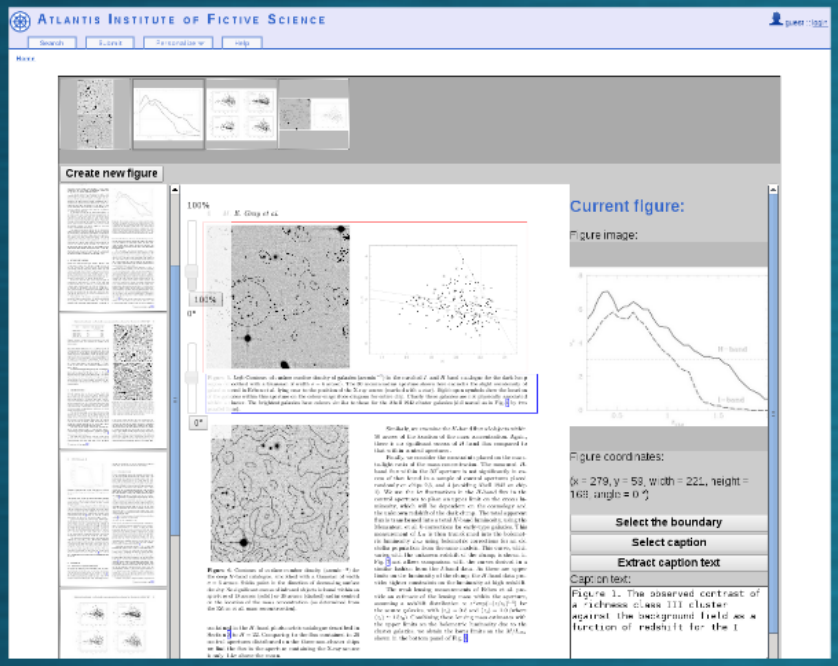
- Distributed execution on a (dynamic) cluster of machines
- Task-level failover
  - Timeouts/error handling
  - Resubmission
- No installation required

# Extractor Merger

- Accepts the output of PDF and LaTeX extractors
- Detects the same figures in both outputs
- Merges the annotation to create a single annotation which is better than separate sources

Code at: https://github.com/ppiotr/Invenio/tree/docextract
( to be rebased, squashed, tested with the new Invenio)

The Manual extractor on the Invenio demo site




- Allows to browse the publication page by page
- Allows to select figures and captions
- Attempts to automatically extract text of captions
- Rotations and scaling allow to correctly extract figures from scanned document

## The Mechanism of Approvals

- Changes are saved to the waiting list
- The operator can view the saved changes, edit them an apply to the main database

## TODO

Status: "pre-alpha" - it is possible to use it with arbitrary records, some functionalities are missing


https://github.com/ppiotr/Invenio/tree/manualplotsextractor

Access: /recor/number/pageimages/
example: http://localhost/record/10/pageimages/

# Manual Extractor

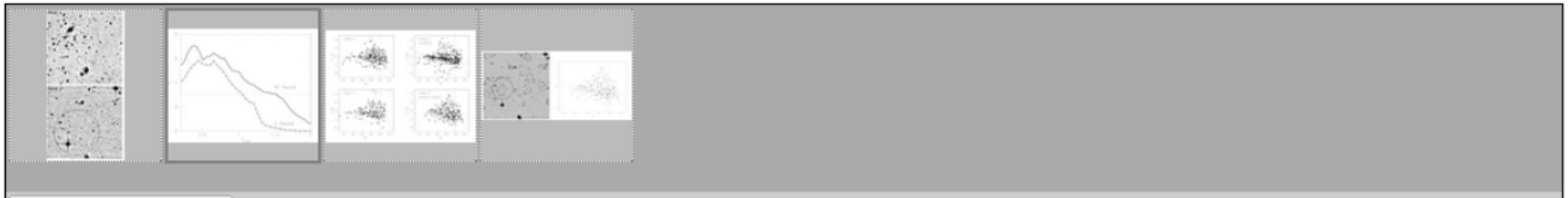
# The manual extractor on the Invenio demo site

ATLANTIS INSTITUTE OF FICTIVE SCIENCE

guest :: login

Search | Submit | Personalize | Help

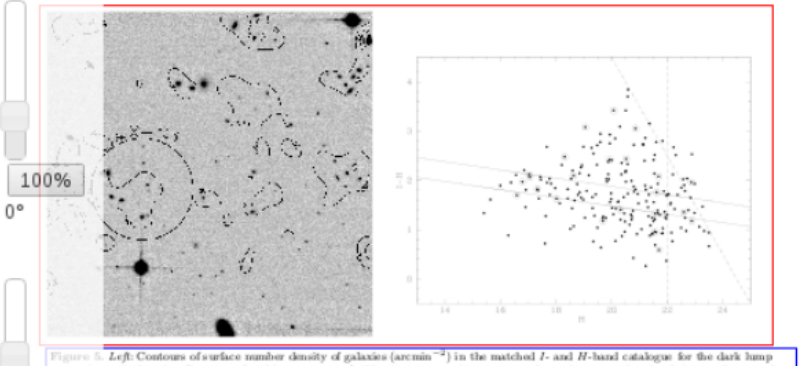
Home



Create new figure

100%

0°



6 M. E. Gray et al.

Figure 5. *Left*: Contours of surface number density of galaxies (arcmin$^{-2}$) in the matched $I$- and $H$-band catalogue for the dark lump region (smoothed with a Gaussian of width $\sigma = 6$ arcsec). The 30 arcmin radius aperture shown here encircles the slight overdensity of galaxies noted in Erben et al. lying near to the position of the X-ray source (marked with a star). *Right*: open symbols show the location of the galaxies within this aperture on the colour-magnitude diagram for entire chip. Clearly these galaxies are not physically associated within the cluster. The brightest galaxies have colours similar to those for the Abell 1942 cluster galaxies (delineated as in Fig. 4 by two parallel lines).

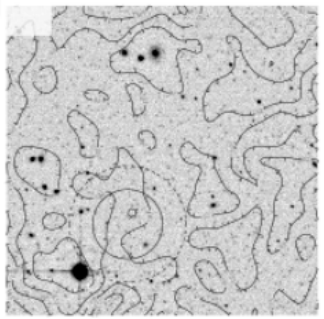


**Figure 6.** Contours of surface number density (arcmin$^{-2}$) for the deep $H$-band catalogue, smoothed with a Gaussian of width $\sigma = 6$ arcsec. Sticks point in the direction of decreasing surface density. No significant excess of infrared objects is found within an aperture of 18 arcsec (solid) or 36 arcsec (dashed) radius centred on the location of the mass concentration (as determined from the Erben et al. mass reconstruction).

contained in the $H$-band photometric catalogue described in Section 4 to $H = 22$. Comparing to the flux contained in 26 control apertures distributed on the three non-cluster chips we find the flux in the aperture containing the X-ray source is only $1.4\sigma$ above the mean.

Similarly, we examine the $H$-band flux of objects within 50 arcsec of the location of the mass concentration. Again, there is no significant excess of $H$-band flux compared to that within control apertures.

Finally, we consider the constraints placed on the mass-to-light ratio of the mass concentration. The measured $H$-band flux within the 50″ aperture is not significantly in excess of that found in a sample of control apertures placed randomly on chips 2,3, and 4 (avoiding Abell 1942 on chip 1). We use the $1\sigma$ fluctuations in the $H$-band flux in the control apertures to place an upper limit on the excess luminosity, which will be dependent on the cosmology and the unknown redshift of the dark clump. The total apparent flux is transformed into a total $H$-band luminosity, using the Menanteau et al. $k$-corrections for early-type galaxies. This measurement of $L_H$ is then transformed into the bolometric luminosity $L_{bol}$ using bolometric corrections for an old stellar population from the same models. This curve, which varies with the unknown redshift of the clump, is shown in Fig. 7 and allows comparison with the curves derived in a similar fashion from the $I$-band data. As these are upper limits on the luminosity of the clump, the $H$-band data provides tighter constraints on the luminosity at high redshift.

The weak lensing measurements of Erben et al. provide an estimate of the lensing mass within the aperture, assuming a redshift distribution $\propto z^2 \exp[-(z/z_0)^{1.5}]$ for the source galaxies, with $\langle z_s \rangle = 0.8$ and $\langle z_s \rangle = 1.0$ (where $\langle z_s \rangle \simeq 1.5 z_0$). Combining these lensing mass estimates with the upper limits on the bolometric luminosity due to the cluster galaxies, we obtain the lower limits on the $M/L_{bol}$ shown in the bottom panel of Fig. 7.

## Current figure:

Figure image:

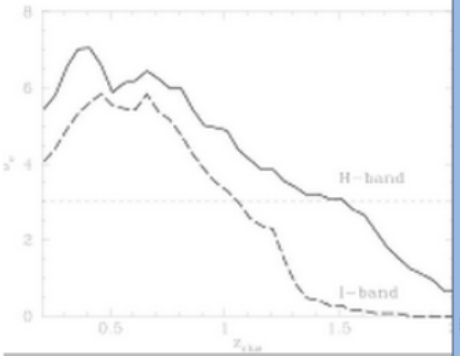



Figure coordinates:

(x = 279, y = 59, width = 221, height = 169, angle = 0 °)

**Select the boundary**

**Select caption**

**Extract caption text**

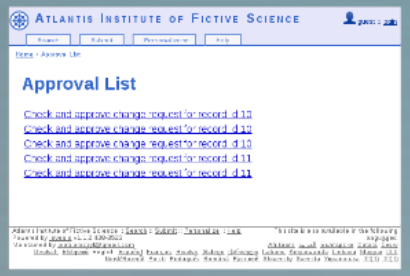
Caption text:

Figure 1. The observed contrast of a richness class III cluster against the background field as a function of redshift for the I

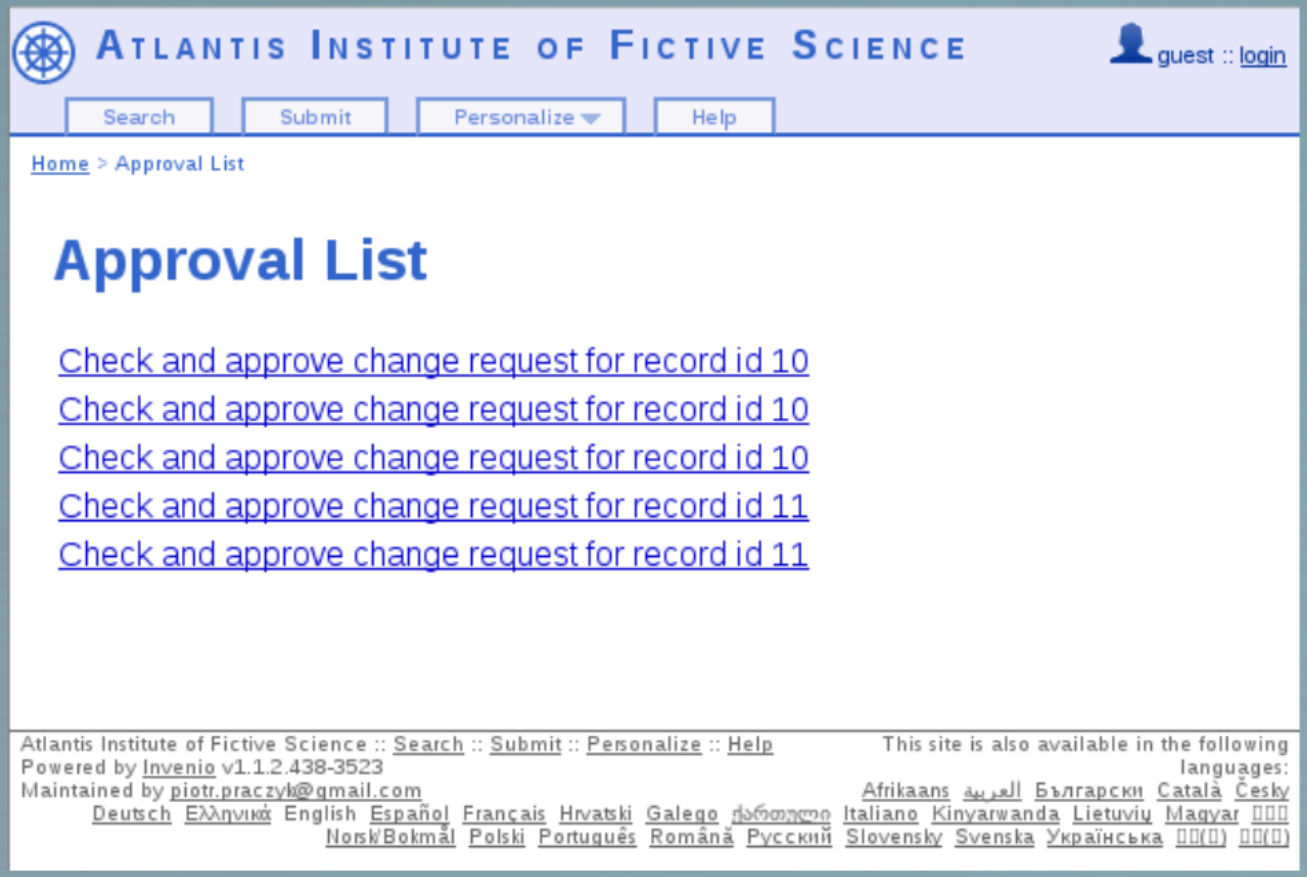
Prezi

# The Mechanism of Approvals

- Changes are saved to the waiting list
- The operator can view the saved changes, edit them an apply to the main database




The very basic approval interface




The very basic approval interface


# TODO

## Status: "pre-alpha" - it is possible to use it with arbitrary records, some functionalities are missing

### Metadata-driven extractor

The module processing the user-generated metadata and generating images + metadata in the upload-friendly format

(1-2 days of work, simple modification of the PDF plots extractor)

### Integration with crowdsourcing workflows

At present everyone has access to the extractor and to the approval interface.

Approval should be limited to the operators and authors (some type of BibAuthorId - like authentication ? )

### Figures extracted from LaTeX

Some figures are not annotated with their location metadata which prevents them from being visualised in the interface. At present they are completely ignored and overriden with the results of the manual selection

LaTeX-exactor should be extended with a module extending the extracted figures with location-metadata

### Extension with semantic-annotation tools

At present, figures can be manually annotated only with their environment metadata (placing it in the context of a publication).

The manual extractor could be used to select elements of the semantic annotation.

Feedback loop could be used to influence the results of the automatic semantics extraction

# Metadata-driven extractor

The module processing the user-generated metadata and generating images + metadata in the upload-friendly format

(1-2 days of work, simple modification of the PDF plots extractor)

# Integration with crowdsourcing workflows

At present everyone has access to the extractor and to the approval interface.

Approval should be limited to the operators and authors (some type of BibAuthorId - like authentication ? )

# Figures extracted from LaTeX

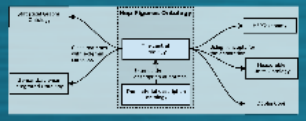
Some figures are not annotated with their location metadata which prevents them from being visualised in the interface. At present they are completely ignored and overriden with the results of the manual selection

LaTeX-exactor should be extended with a module extending the extracted figures with location-metadata

# Annotation

## HEP Figures Ontology

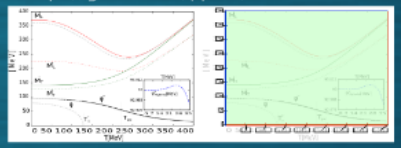
- The RDF data model for describing figures (the publication-level description and the internal structure of figures)
- Abstract model - mapped on top of the BibDocFile and BibRecord structures



## Semantics Extraction

- Analyses the vector graphics of the extracted figures
- Detects the coordinate systems (together with accompanying objects)
- Generated the annotation according to HFO

Code at: https://github.com/ppiotr/PlotFeatureExtractor



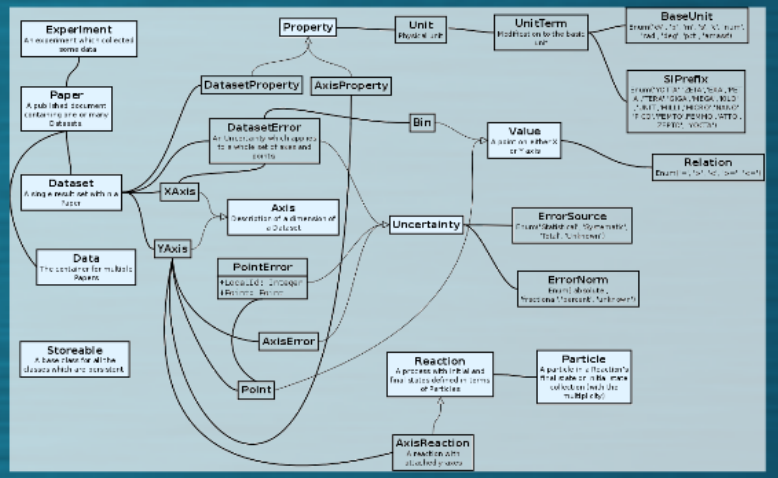
## BibDocFile data model

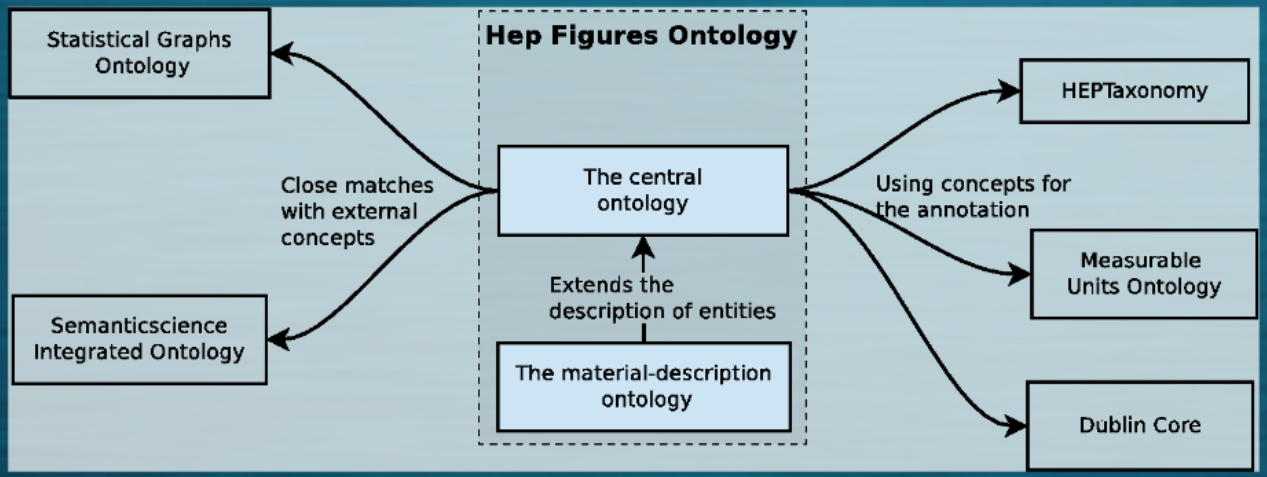
- Implemented
- Integrated in Master
- Not deployed in the production system

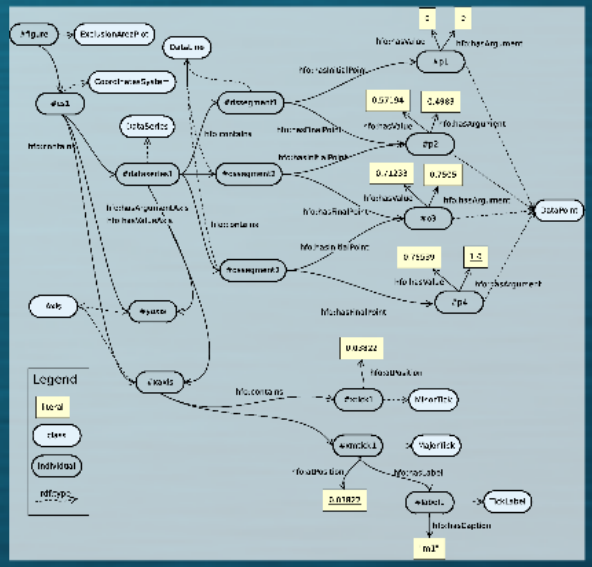
The state of the deployment/implementation of different components

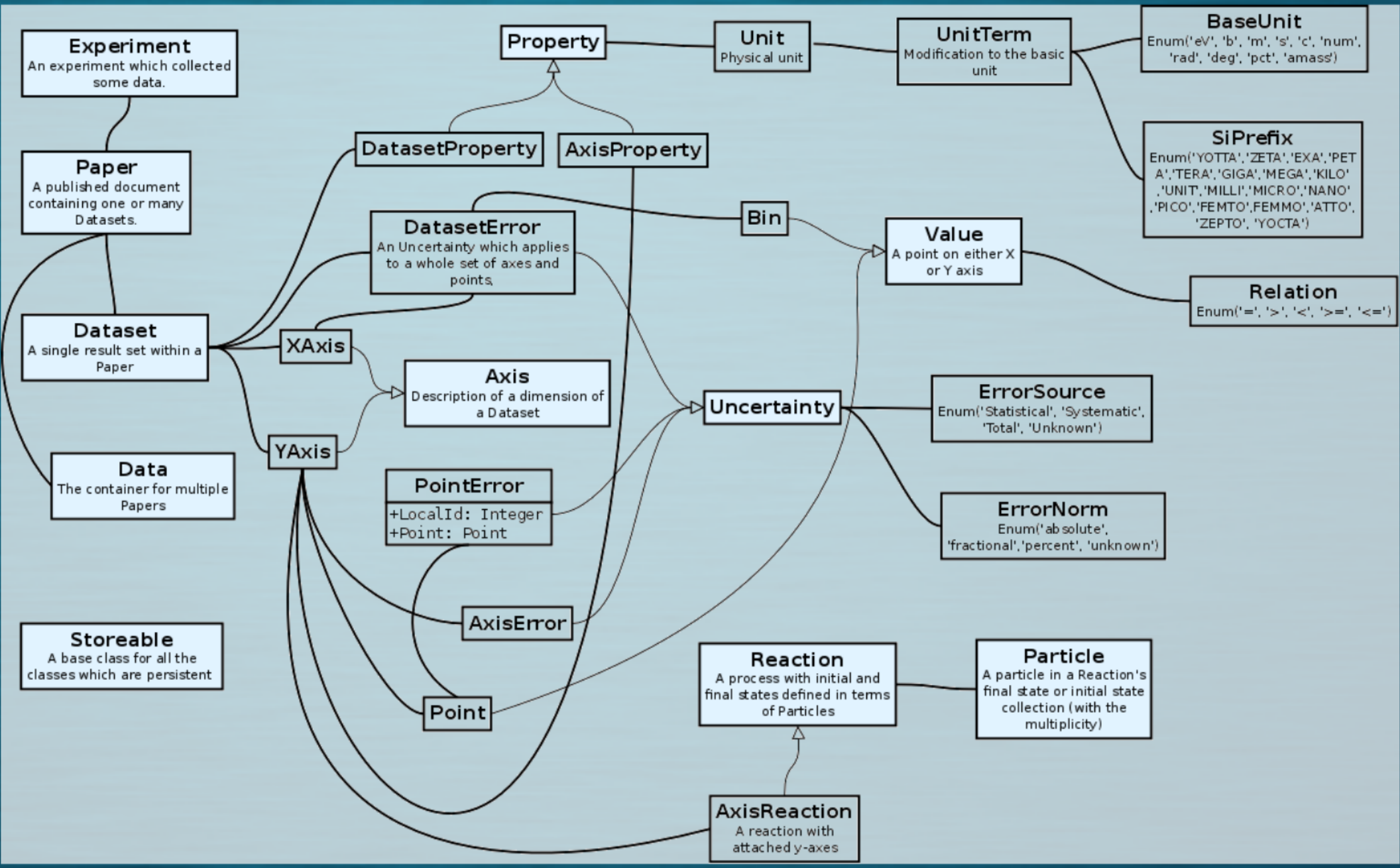
# HEP Figures Ontology

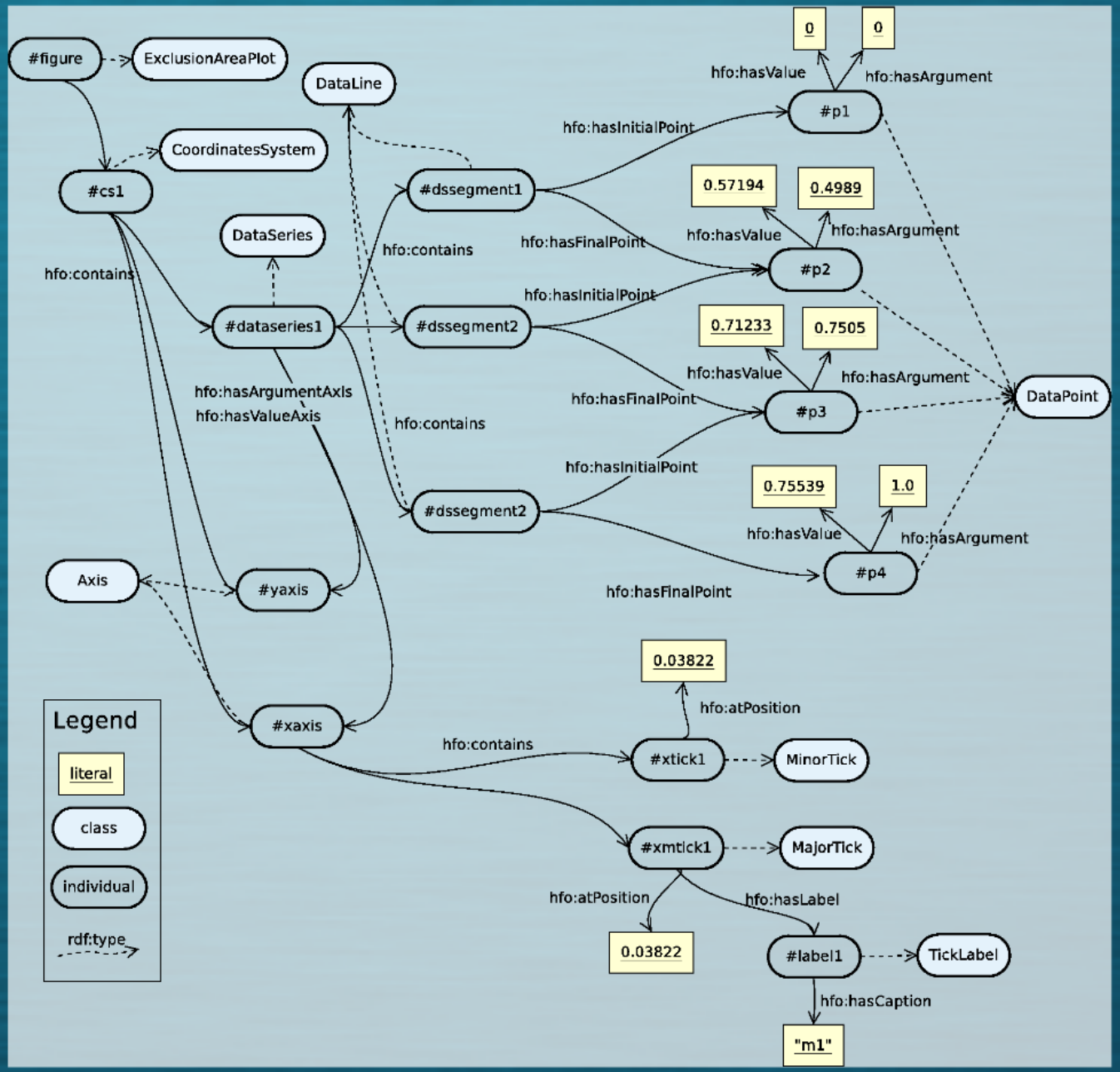
- The RDF data model for describing figures (the publication-level description and the internal structure of figures)
- Abstract model - mapped on top of the BibDocFile and BibRecord structures



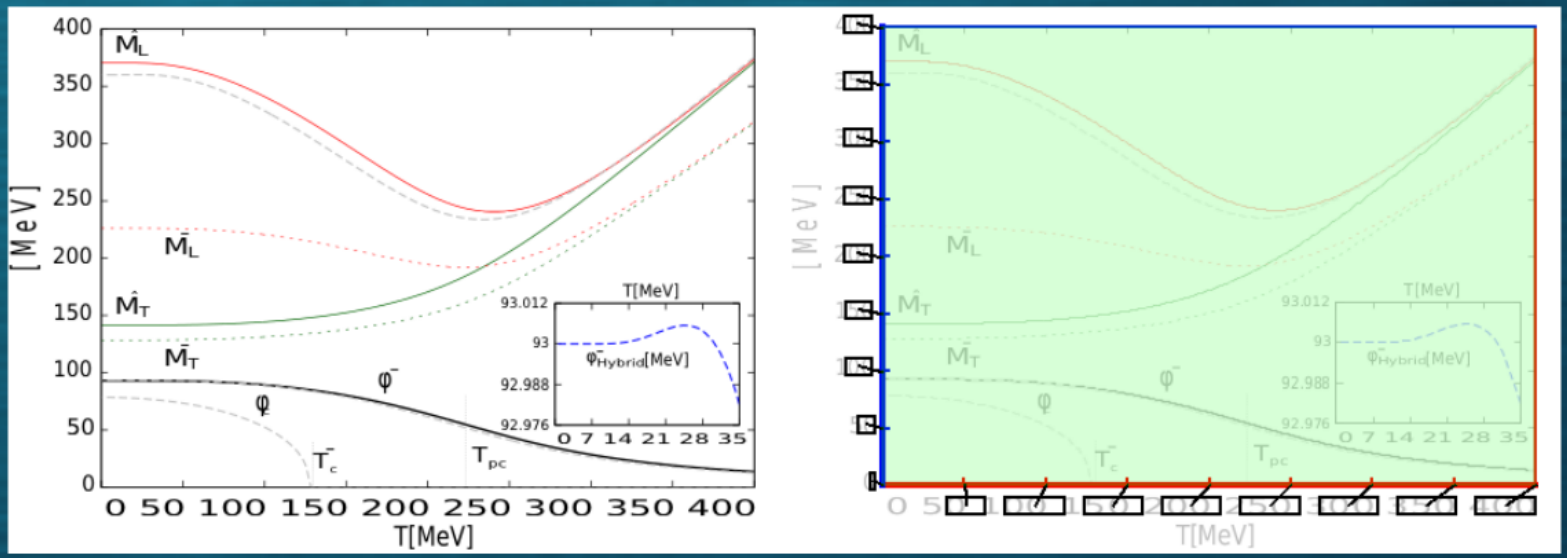

Experiment
An experiment which collected some data.
Paper
A published document containing one or many Datasets.
Dataset
A single result set within a Paper
Data
The container for multiple Papers
Storeable
A base class for all the classes which are persistent
Property
Unit
Physical unit
UnitTerm
Modification to the basic unit
BaseUnit
Enum('eV', 'b', 'm', 's', 'c', 'num', 'rad', 'deg', 'pct', 'amass')
SiPrefix
Enum('YOTTA','ZETA','EXA','PETA','TERA','GIGA','MEGA','KILO','UNIT','MILLI','MICRO','NANO','PICO','FEMTO','FEMMO','ATTO','ZEPTO', 'YOCTA')
DatasetProperty
AxisProperty
DatasetError
An Uncertainty which applies to a whole set of axes and points
Bin
Value
A point on either X or Y axis
Relation
Enum('=', '>', '<', '>=', '<=')
XAxis
YAxis
Axis
Description of a dimension of a Dataset
Uncertainty
ErrorSource
Enum('Statistical', 'Systematic', 'Total', 'Unknown')
ErrorNorm
Enum('absolute', 'fractional','percent', 'unknown')
PointError
+LocalId: Integer
+Point: Point
AxisError
Point
Reaction
A process with initial and final states defined in terms of Particles
Particle
A particle in a Reaction's final state or initial state collection (with the multiplicity)
AxisReaction
A reaction with attached y-axes


#figure
ExclusionAreaPlot
DataLine
CoordinatesSystem
#cs1
DataSeries
hfo:contains
#dataseries1
hfo:hasArgumentAxis
hfo:hasValueAxis
#dssegment1
hfo:contains
#dssegment2
hfo:contains
#dssegment2
0
0
hfo:hasValue
hfo:hasArgument
#p1
hfo:hasInitialPoint
0.57194
0.4989
hfo:hasValue
hfo:hasArgument
#p2
hfo:hasFinalPoint
hfo:hasInitialPoint
0.71233
0.7505
hfo:hasValue
hfo:hasArgument
#p3
hfo:hasFinalPoint
hfo:hasInitialPoint
0.75539
1.0
hfo:hasValue
hfo:hasArgument
#p4
hfo:hasFinalPoint
DataPoint
Axis
#yaxis
#xaxis
0.03822
hfo:atPosition
hfo:contains
#xtick1
MinorTick
#xmtick1
MajorTick
hfo:atPosition
hfo:hasLabel
0.03822
#label1
TickLabel
hfo:hasCaption
"m1"
Legend
literal
class
individual
rdf:type

# Semantics Extraction

- Analyses the vector graphics of the extracted figures
- Detects the coordinate systems (together with accompanying objects)
- Generated the annotation according to HFO

Code at: https://github.com/ppiotr/PlotFeatureExtractor



# Extension with semantic-annotation tools

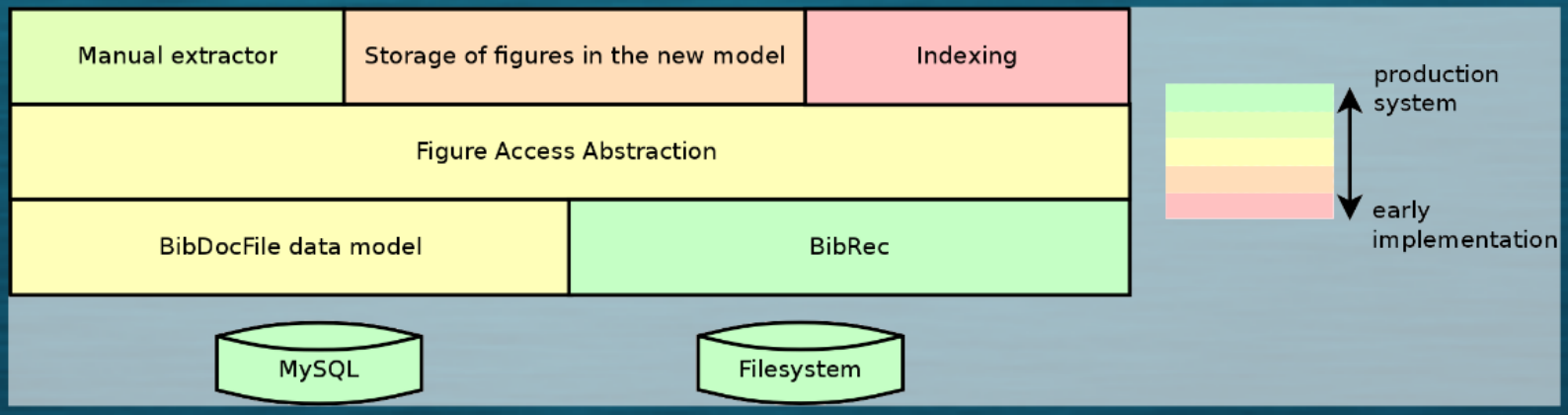
At present, figures can be manually annotated only with their environment metadata (placing it in the context of a publication).

The manual extractor could be used to select elements of the semantic annotation.

Feedback loop could be used to influence the results of the automatic semantics extraction

# BibDocFile data model

- Implemented
- Integrated in Master
- Not deployed in the production system




The state of the deployment/implementation of different components

# Access

## SPARQL searching prototype

- A Prototype implementation in Java (Jena) for testing the SPARQL queries on a corpus of automatically extracted figures
- in INSPIRE: Parts of the annotation could be indexed using internal search engine/ we could use external semanic search engine

Code at: https://github.com/ppiotr/InvenioSemantics

## HEPData

- The case similar to figures (datasets stored as separate records)
- Figures can link to records (implemented, currently not used as links have to be added manually at the moment)

# SPARQL searching prototype

- A Prototype implementation in Java (Jena) for testing the SPARQL queries on a corpus of automatically extracted figures
- in INSPIRE: Parts of the annotation could be indexed using internal search engine/ we could use external semanic search engine

Code at: https://github.com/ppiotr/InvenioSemantics

# HEPData

- The case similar to figures (datasets stored as separate records)
- Figures can link to records (implemented, currently not used as links have to be added manually at the moment)