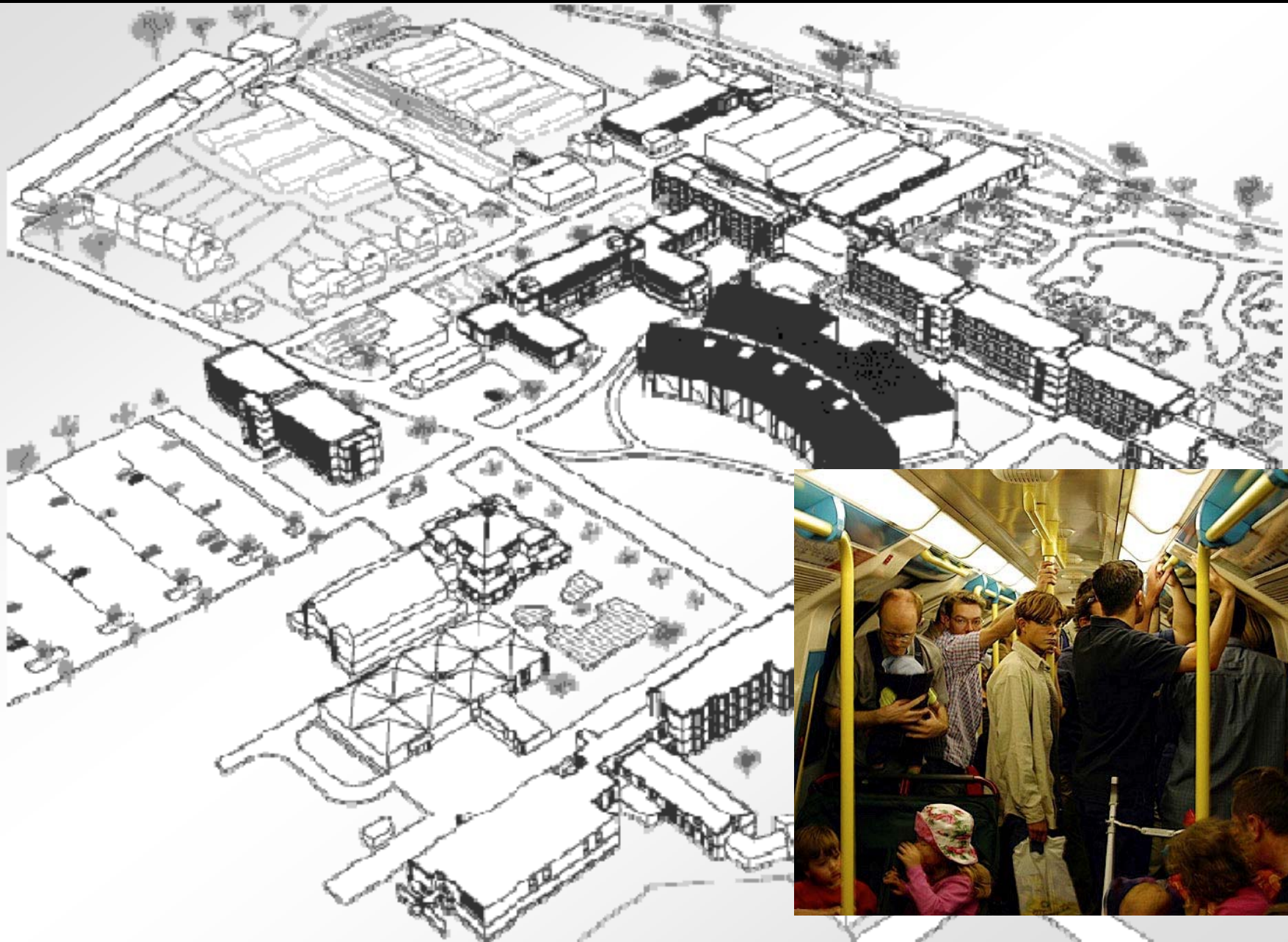
# The ONDEX Framework: Uniting Concept-based Data Integration, Text Mining, Biological Homology Searches and Data Analysis

*Jacob Köhler*

*Rothamsted Research*

# RRes

# RRes



# Credits

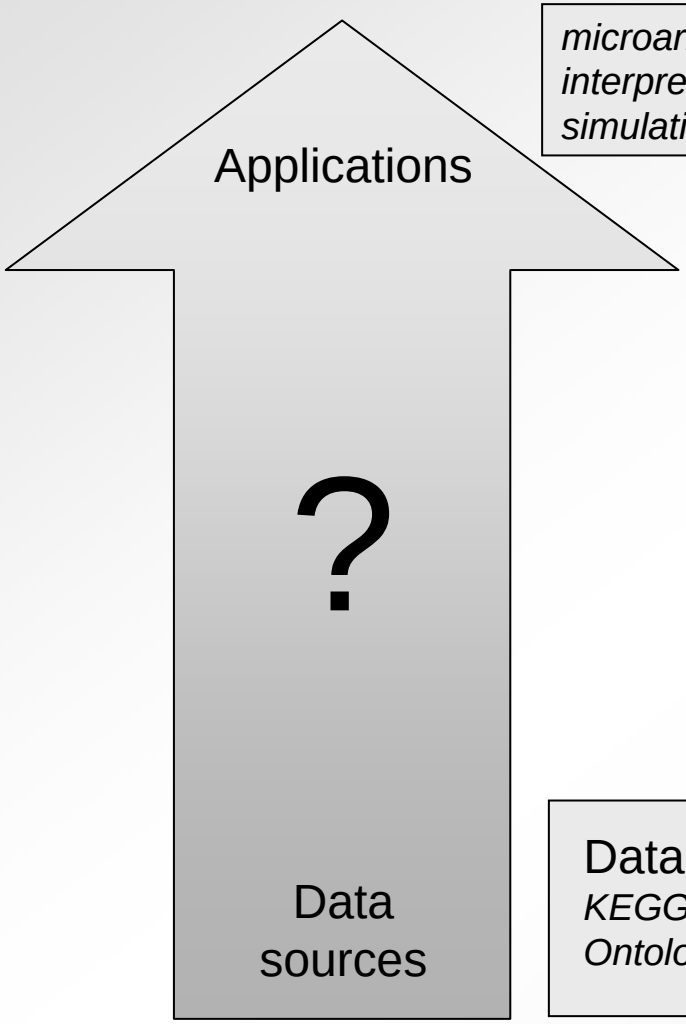
Jan Baumbach (Rothamsted & Bielefeld)
Jessica Butz (Rothamsted & Bielefeld)
Ina Kupp (Rothamsted & Koblenz)
Stephan Philippi (Koblenz)
Chris Rawlings (Rothamsted)
Alexander Rüegg (Bielefeld)
Andre Skusa (Bielefeld)
Michael Specht (Rothamsted & Bielefeld)
Jan Taubert (Rothamsted & Bielefeld)
Paul Verrier (Rothamsted)
Rainer Winnenburg (Rothamsted & Bielefeld)

Rothamsted Research, Harpenden, Hertfordshire, UK.
University of Bielefeld, Germany
University of Koblenz, Germany

# Overview

- Motivation and Introduction
- Principles
- ONDEX System
- Applications

# Motivation and introduction




*microarray analysis and interpretation, modelling and simulation, gene annotation*

*Support database curation, Hypothesis generation*

Graph analysis

Sequence analysis

Text mining

Database Integration

Databases – 100s

*KEGG, MetaCyc, Aracyc, Gene Ontology, EC nomenclature, ….*

**Text sources**

*MEDLINE*

*$15x10^6$* abstracts

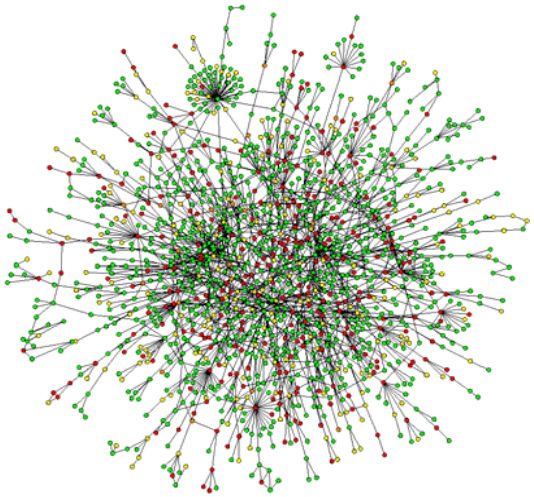
# Motivation and introduction

Combining

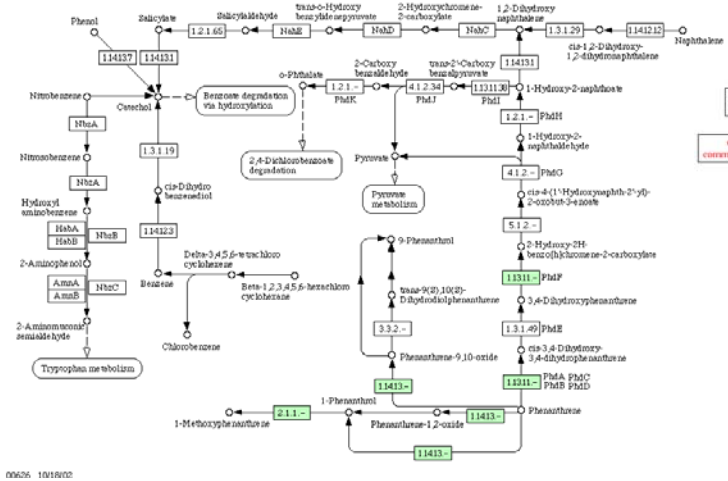
- concept based data integration
- concept based  text mining
- graph analysis/visualisation
- sequence analysis

# Principles

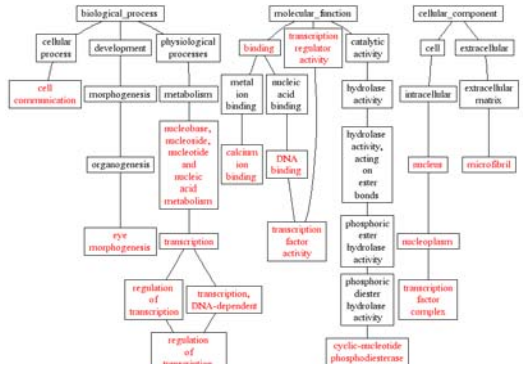
everything is a network…



protein interactions




metabolic pathways

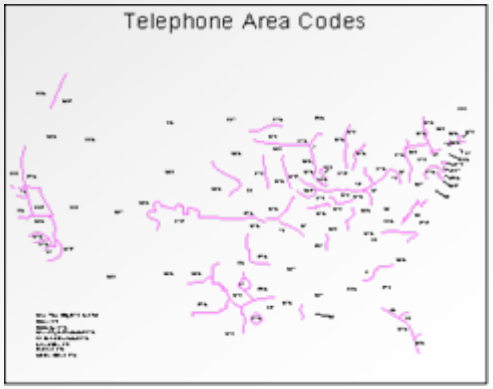


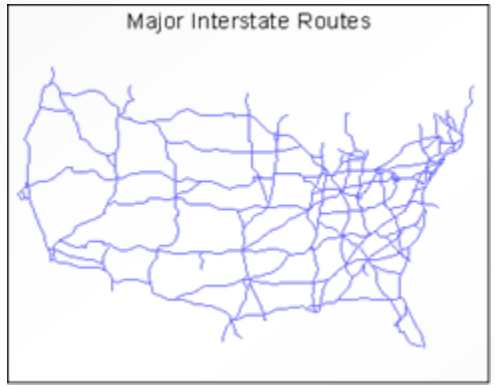

ontologies

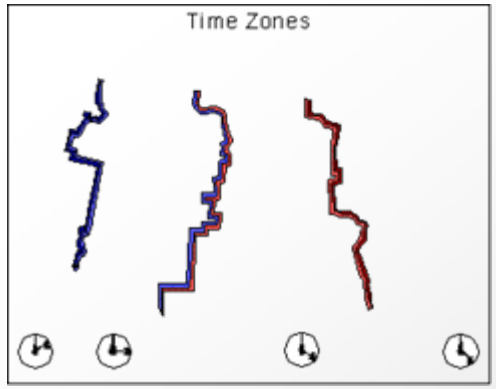
… in which the nodes and edges have different properties

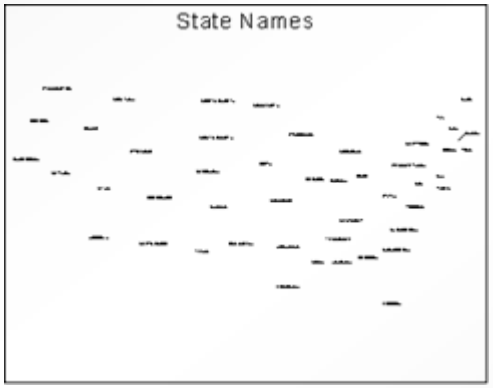
# Principles

Think of it as layers which can separately be added or removed



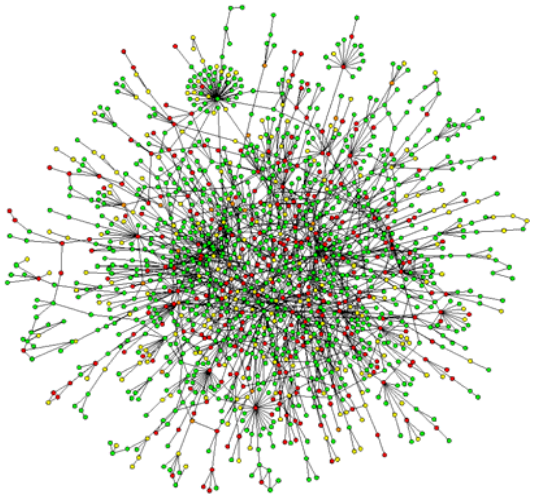



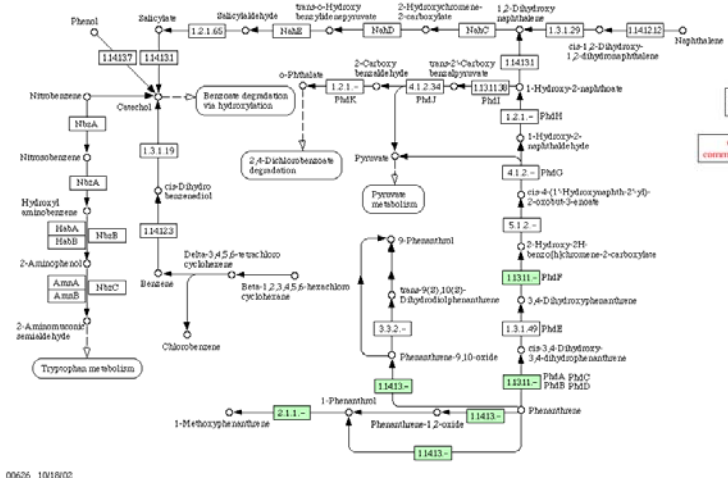


# Principles

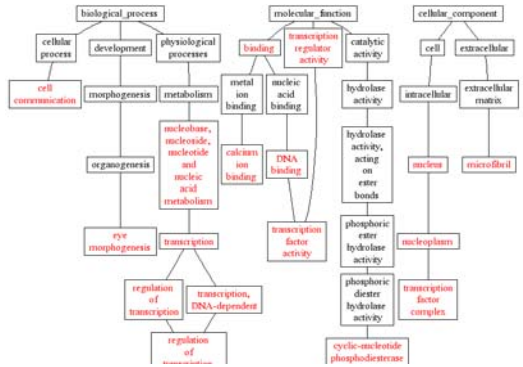
everything is a network…



protein interactions



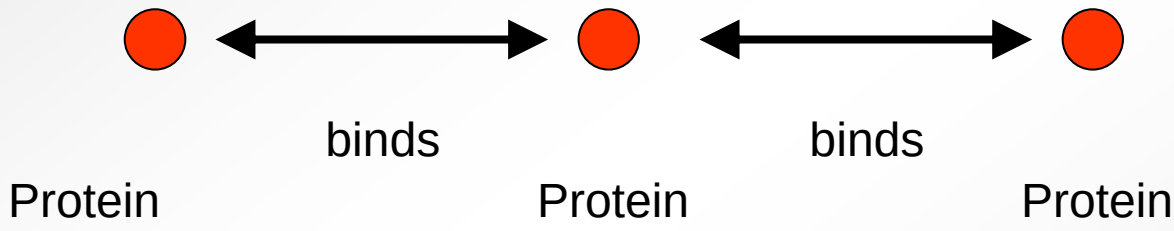

metabolic pathways



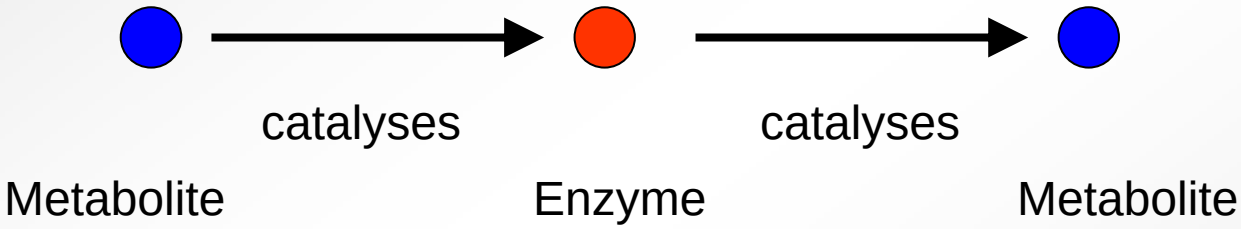

ontologies

… in which the nodes and edges have different properties

# Principles



# Principles



# Principles

## Data Structure

*Integrated ontology* O(C, R, CA, CV, CC, RT, P, ca, cv, cc, rt, id)

- a finite, not empty, distinct set of Concepts C(O)
- a finite, not empty set of Relations: $R(O) \subseteq C(O) \times C(O)$
- a finite set of Concept Accessions CA(O)
- a finite, not empty set of Controlled Vocabularies CV(O)
- a tree consisting of Concept Classes CC(O)
- a tree consisting of Relation Types RT(O)
- the additional properties P(O) of an ontology O' consisting of:
  - a finite set of Concept Names CN(O)
  - a finite set of Sequences SEQ(O)
  - a finite set of Structures STR(O)
- the function ca which assigns concept accessions to concepts
  ca: C(O) → $\{(ca_1 \times \ldots \times ca_n) \mid ca_j \in CA(O)\}$
- the totally defined functions cv, cc, rt that assign CVs, concept classes and relation types to concepts or relations
  cv: $C(O) \cup R(O)$ → CV(O)
  cc: C(O) → CC(O)
  rt: R(O) → RT(O)
- the bijective function id which assigns a unique identifier to every concept and every relation with:
  id: C(O) → **N**
- and the functions def, cn, seq and str that optionally link concept names (terms), definitions, polypeptide or nucleotide sequences and protein structures to concepts:
- def: C(O) → DEF(O)
- cn: C(O) → $\{(cn_1 \times \ldots \times cn_n) \mid cn_j \in CN(O)\}$
- seq: C(O) → $\{(seq_1 \times \ldots \times seq_n) \mid seq_j \in SEQ(O)\}$
- str: C(O) → $\{(str_1 \times \ldots \times str_n) \mid str_j \in STR(O)\}$

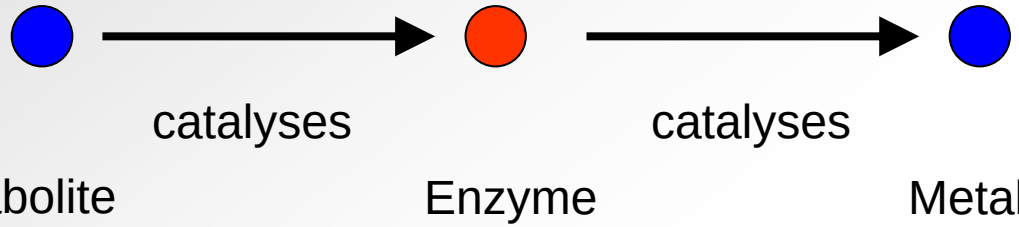
# Principles

## Data Structure

*visible graph* G (O, CO, colour, size, visibility, x, y)

- an integrated ontology, O
- a finite, not empty set of Colours CO(G)
- the functions colour, size, visibility, x and y (coordinates) which affect the way concepts and relations are visualised:
  - colour: C(O) ∪ R(O) → CO(G)
  - size: C(O) → **R**
  - visibility: C(O) ∪ R(O) → {true, false}
  - x: C(O) → **R**
  - y: C(O) → **R**

# Principles




**Concept based data integration and text mining**

| | Db integration | Text mining |
|---|---|---|
| Creating concepts | Db import, conversion, extraction | NER, dictionaries |
| Creating relations | mapping methods, sequence analysis methods | Relation mining |

# Principles

**Mapping methods: graph alignment (not merging!)**

- Only fully automated methods (no manual mapping)

- Use evidence codes to annotate how mappings were generated

- Assign correct semantics (relation type) to the mapping

# Principles

**Mapping methods: graph alignment (not merging!)**

- Import of mapping lists
- Methods based on graph structure (structalign)
- Compare concept names (2syn)
- Sequence analysis:
  Homology, INPARANOID
  (Motif search)
- transitive mapping (trans)

( - protein-protein docking methods)

( - protein-ligand docking methods)

# Principles

**Text mining** 

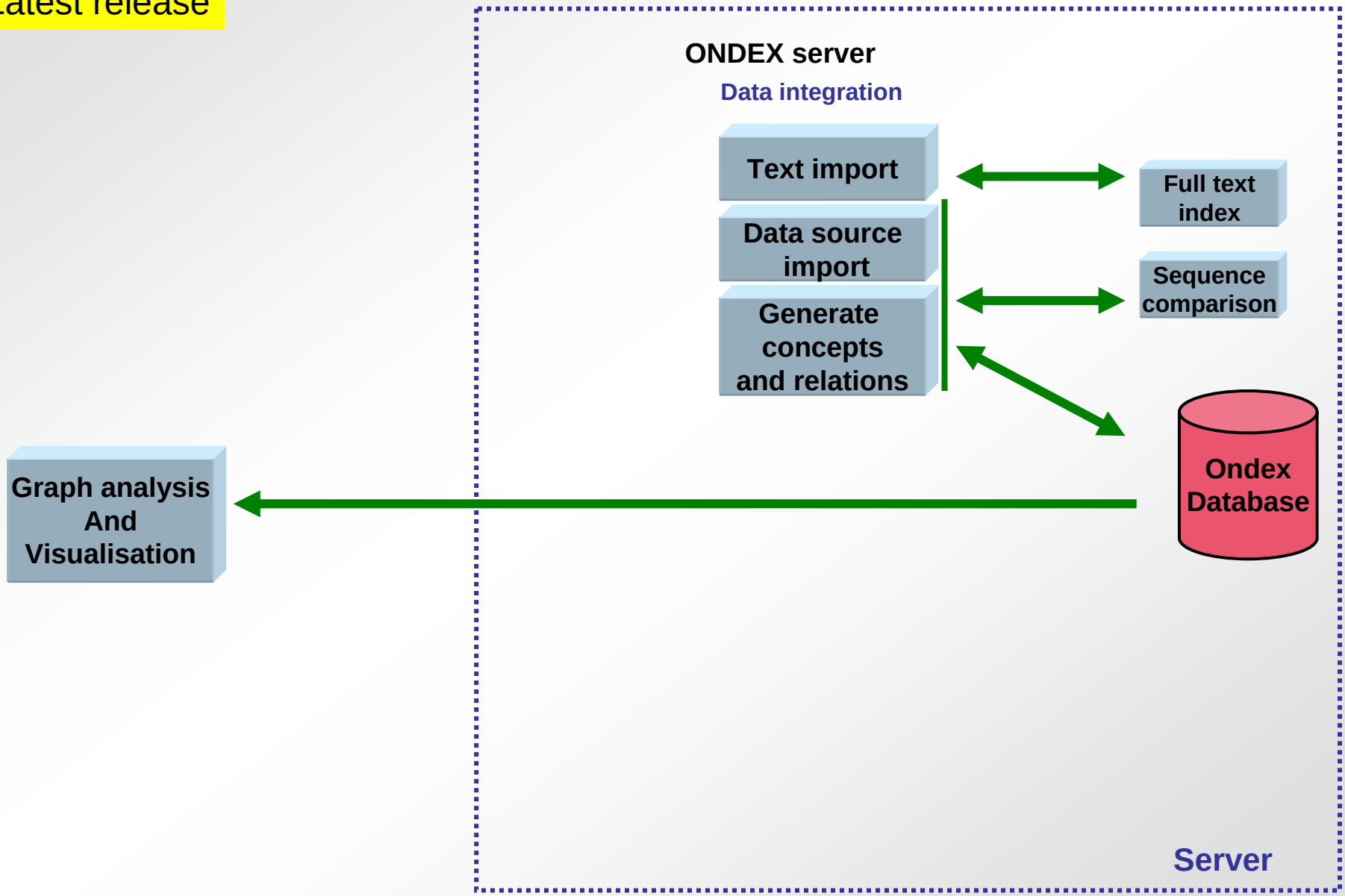
- Concept based approach
- Word stemming, normalisation of concept names, POS tagging
- Concept groups can be defined by
  - a) selected subset of concepts (dictionary approach)
  - b) regular expression
  - c) Planned: other NER methods
- Relation mining
  - a) co-occurrence of concept groups
  - b) planned: deep parsing methods
- Text sources: PubMed

# ONDEX system

Latest release

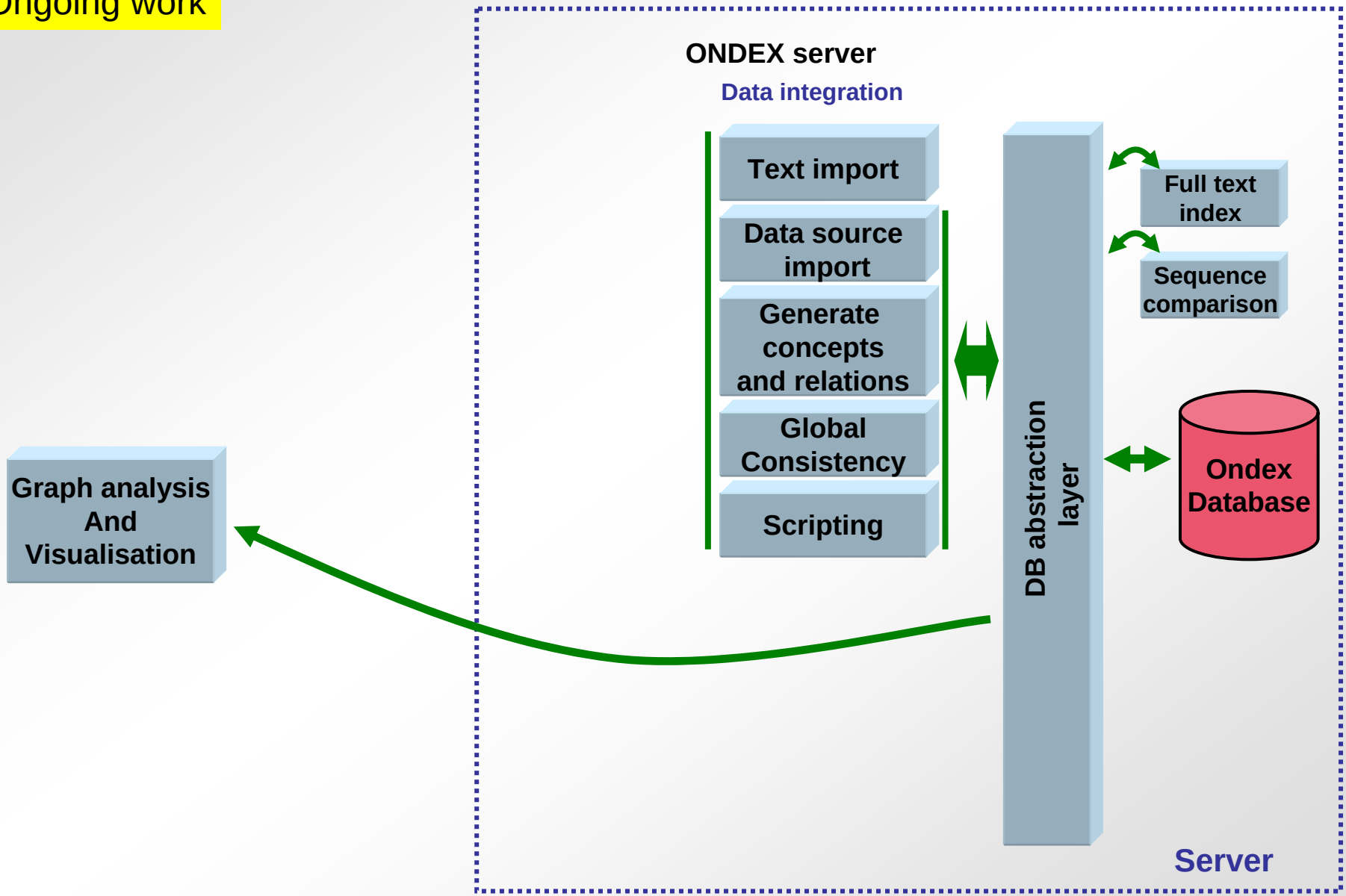
# ONDEX system

Latest release

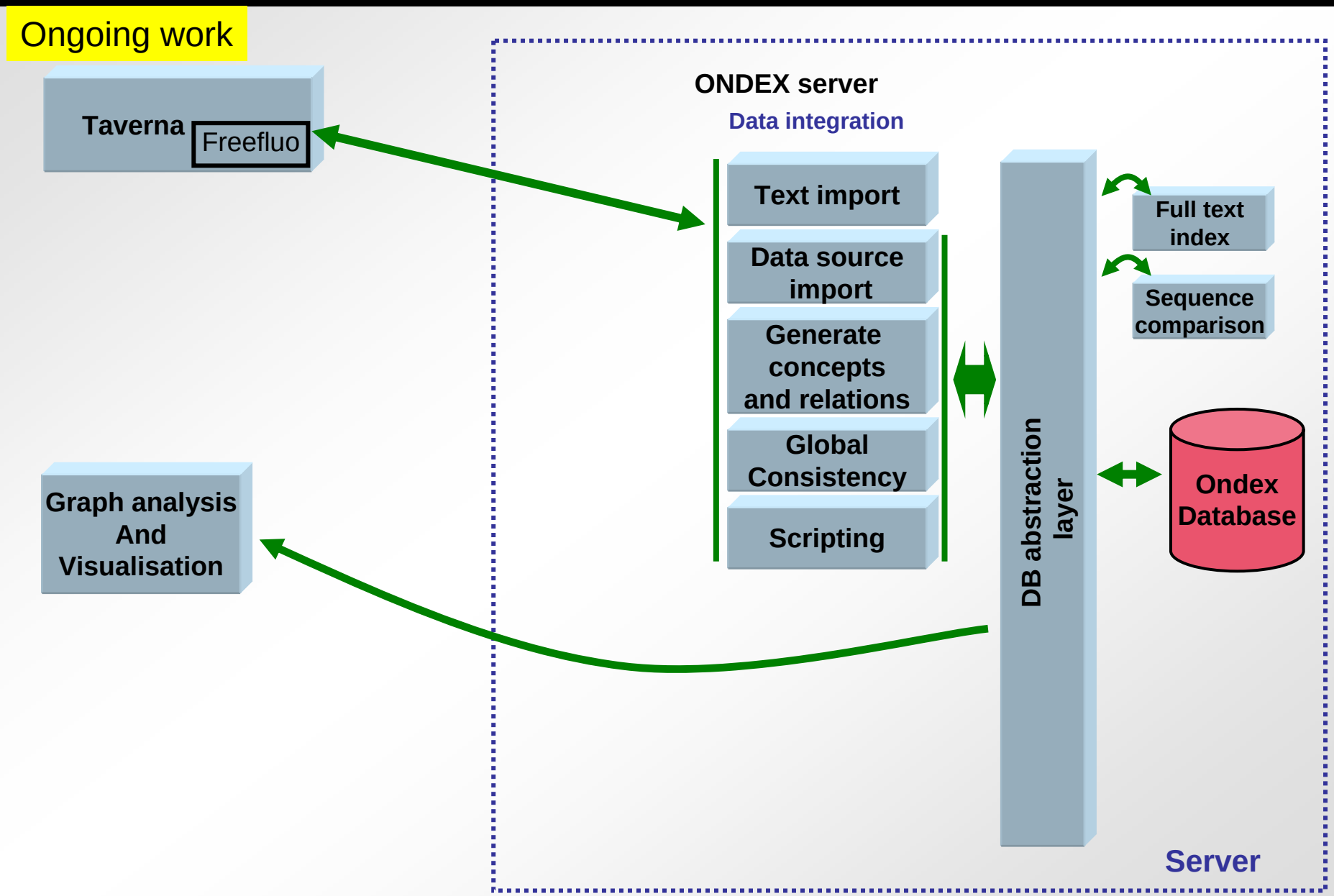


# ONDEX system

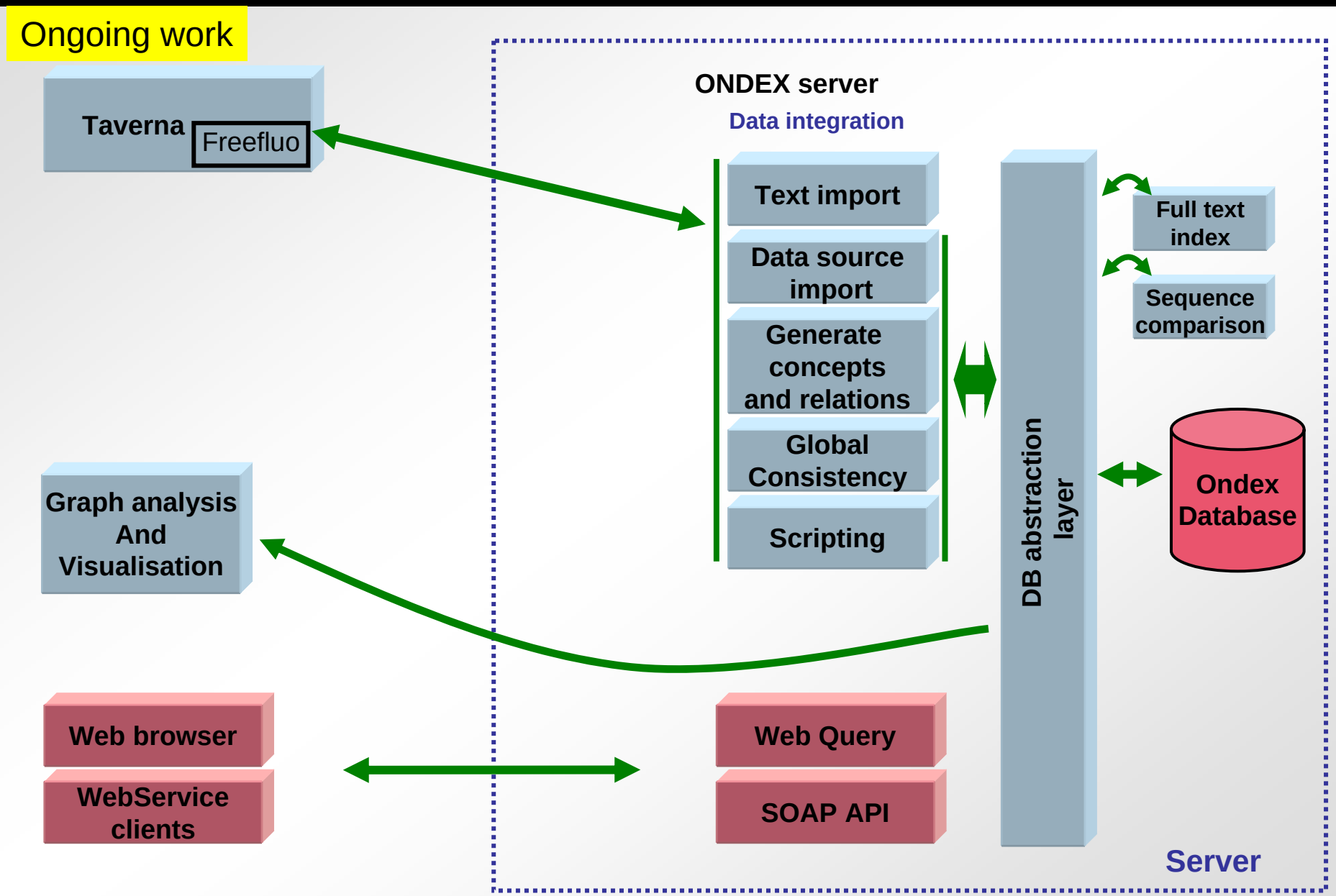
Ongoing work



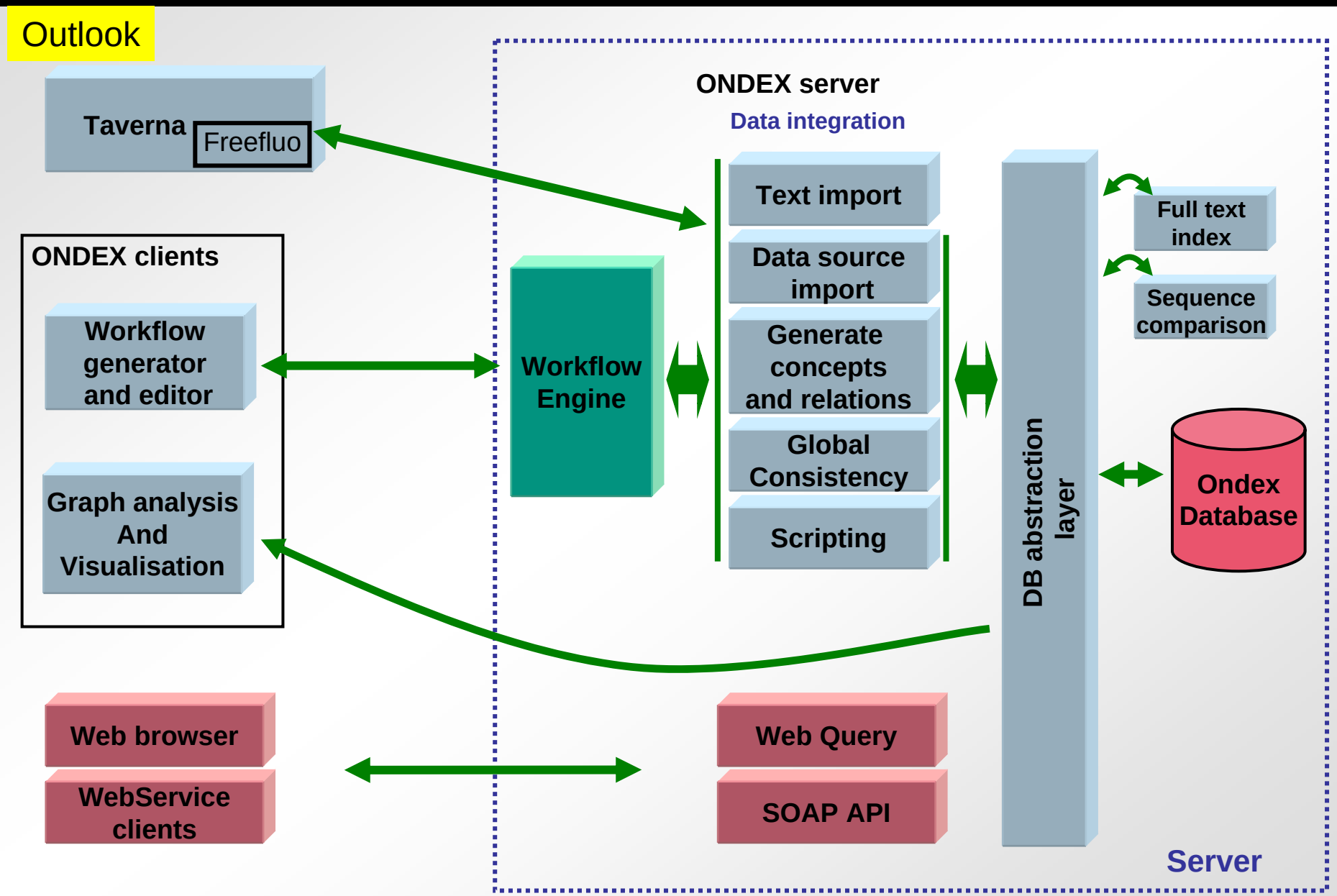
# ONDEX system



# ONDEX system



# ONDEX system



# Applications

- Annotation pipeline

- Microarray analysis

-Text mining to support database curation

- Intelligibility and circularity of terms and definitions in ontologies and taxonomies

- Pathway modelling and simulation

# Applications – microarray analysis

Parani, M., Rudrabhatla, S., Myers, R., Weirich, H., Smith, B., Leaman, D.W. and Goldman, S.L. (2004) Microarray analysis of nitric oxide responsive transcripts in Arabidopsis. Plant Biotechnology Journal, 2, 359-366.

Arabidopsis data with 120 novel genes

→ provided annotation to 71 “novels”

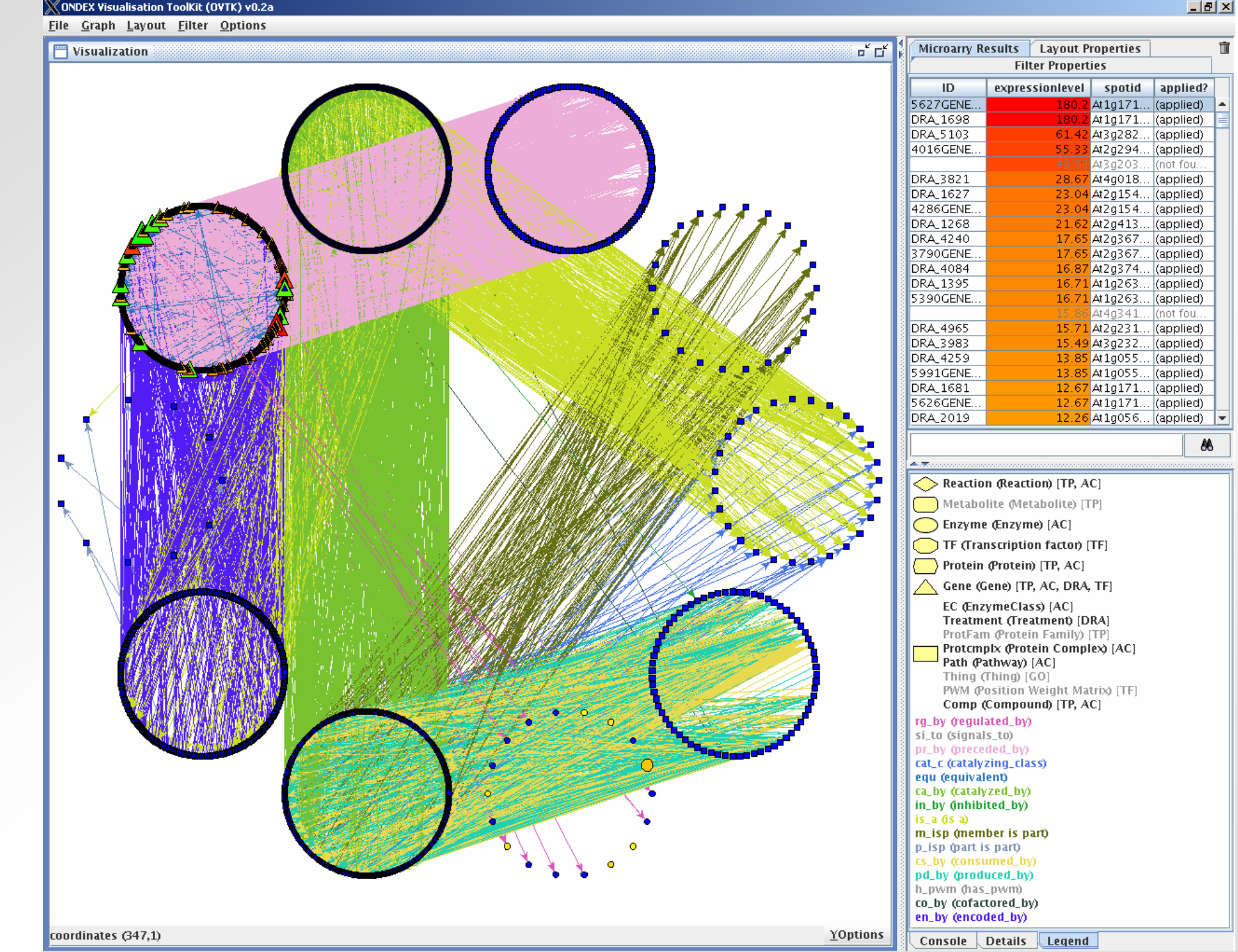
→ lignin biosynthesis

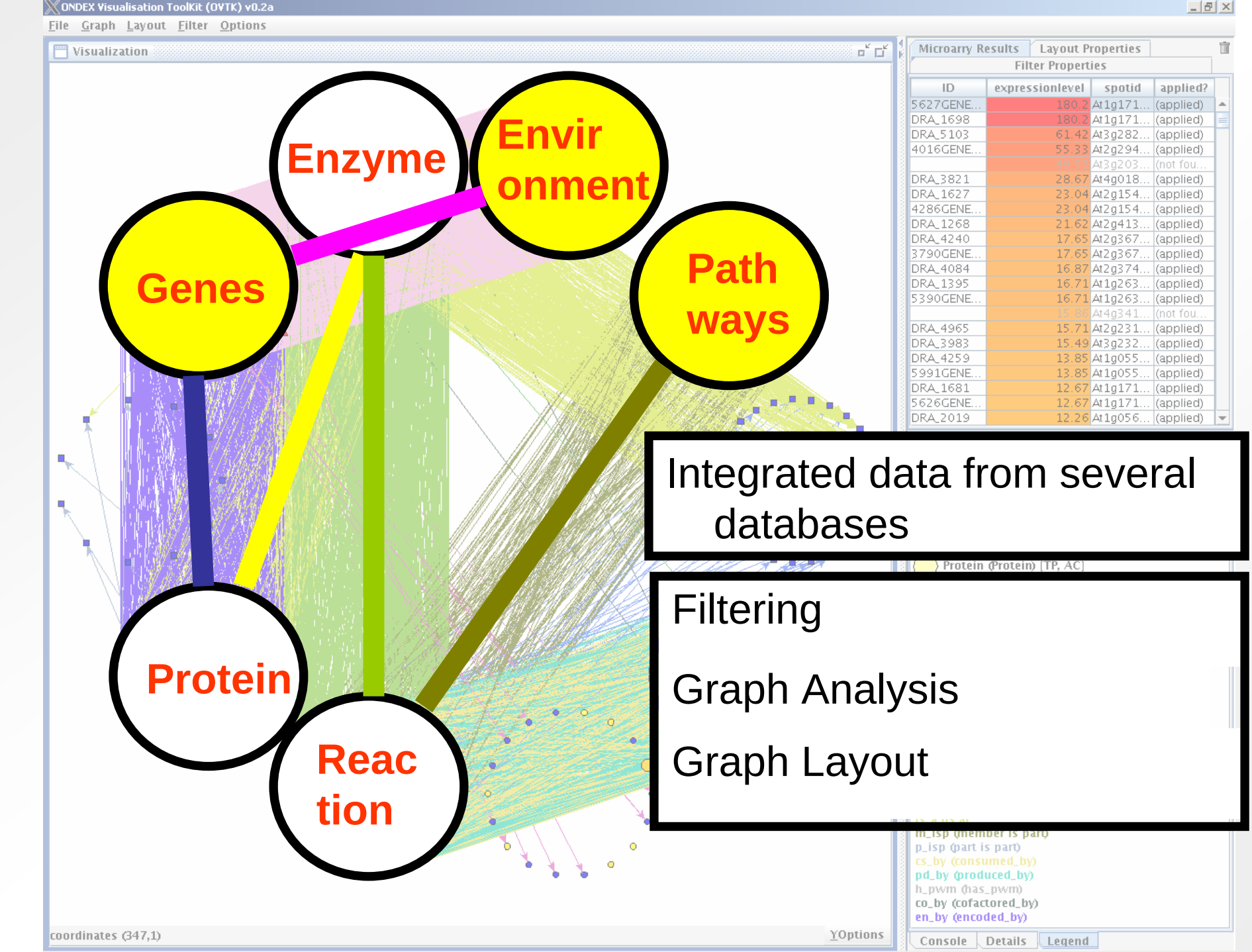
New Observations not in original paper:

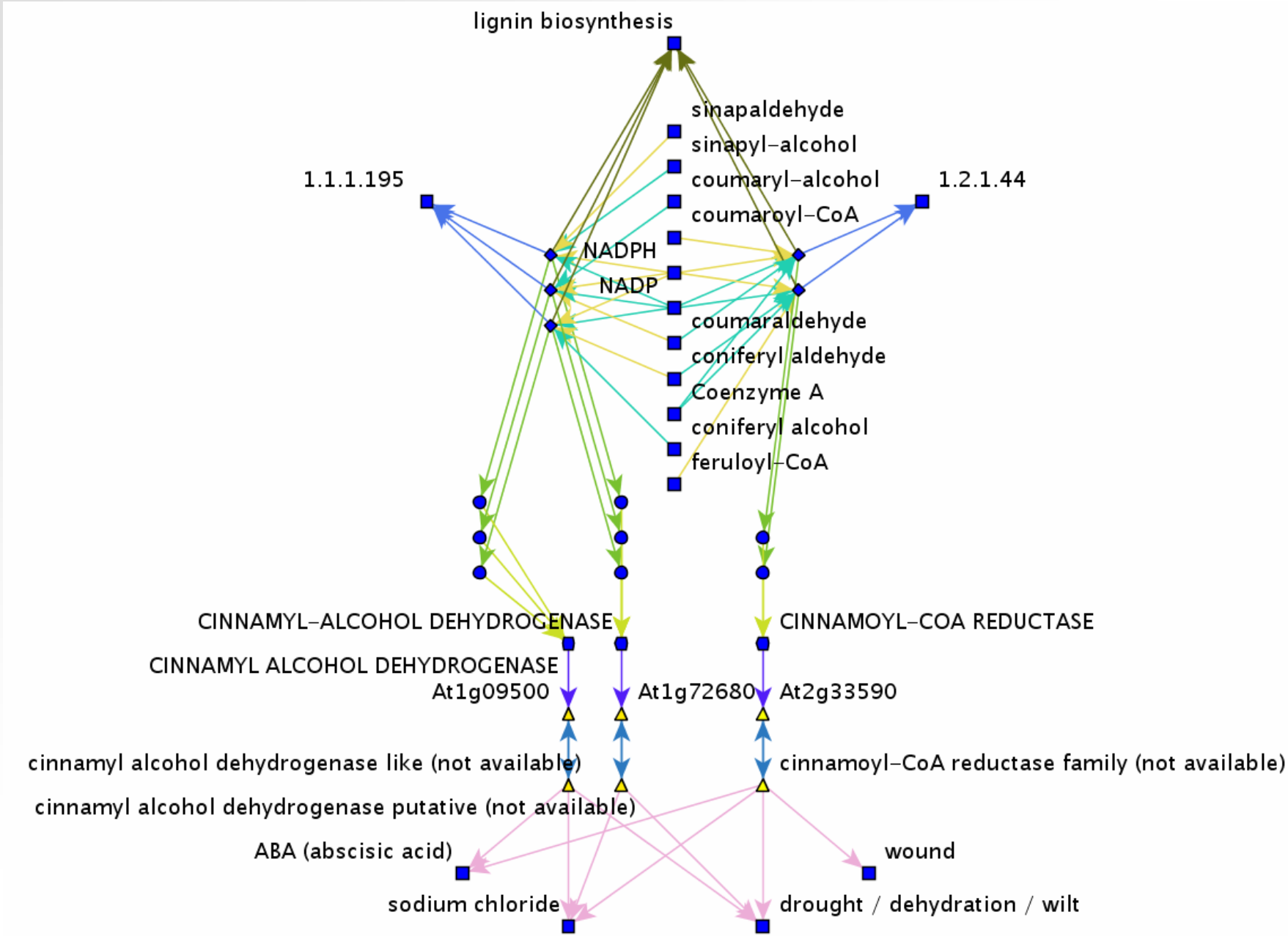
→ Overexpressed transcription factor but no effect on expected gene

→ draught stress

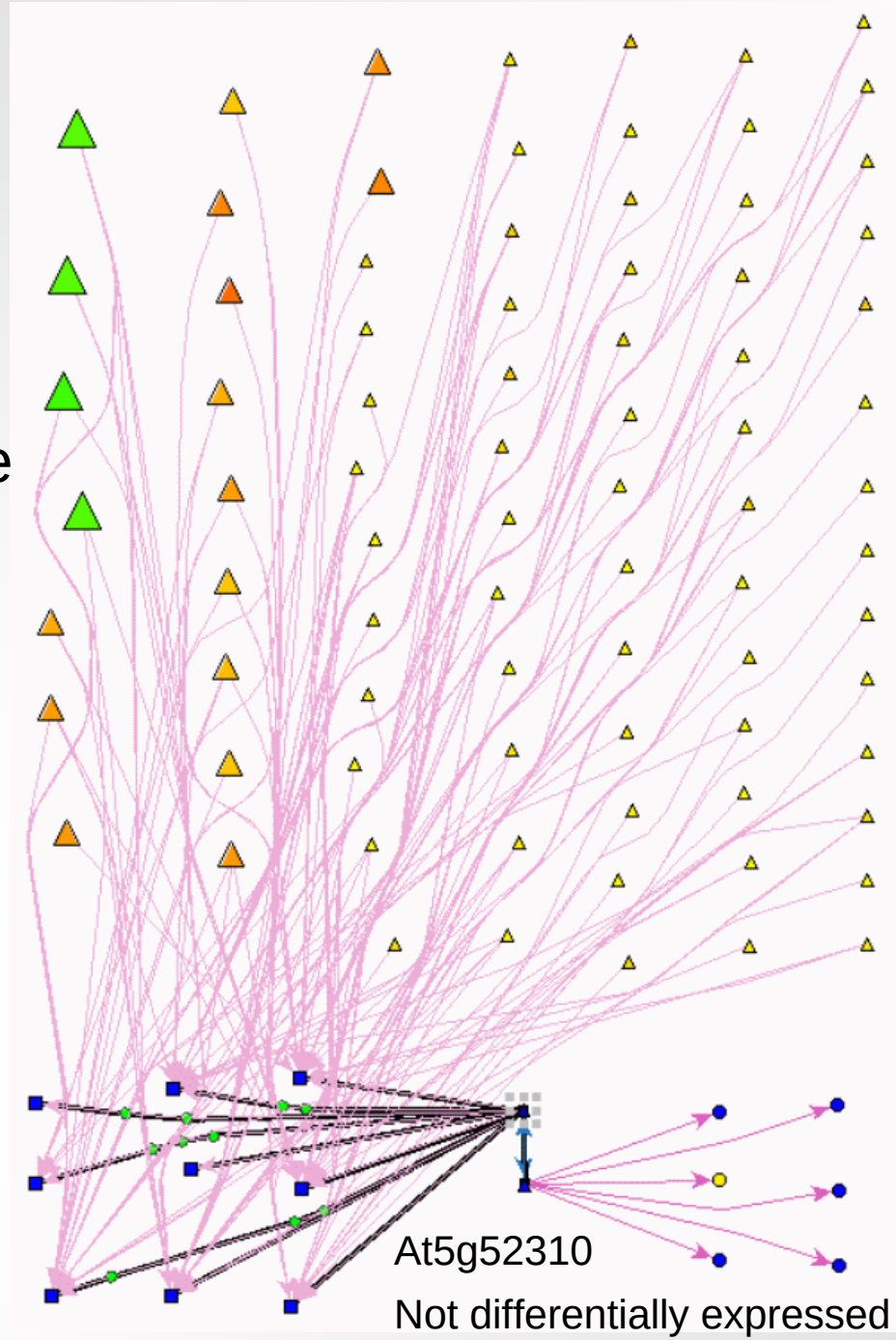
→ jasmonic acid biosynthesis


ONDEX Visualisation ToolKit (OVTK) v0.2a
File Graph Layout Filter Options
Visualization
Microarry Results
Layout Properties
Filter Properties
ID expressionlevel spotid applied?
5627GENE... 180.2 At1g171... (applied)
DRA_1698 180.2 At1g171... (applied)
DRA_5103 61.42 At3g282... (applied)
4016GENE... 55.33 At2g294... (applied)
At3g203... (not fou...
DRA_3821 28.67 At4g018... (applied)
DRA_1627 23.04 At2g154... (applied)
4286GENE... 23.04 At2g154... (applied)
DRA_1268 21.62 At2g413... (applied)
DRA_4240 17.65 At2g367... (applied)
3790GENE... 17.65 At2g367... (applied)
DRA_4084 16.87 At2g374... (applied)
DRA_1395 16.71 At1g263... (applied)
5390GENE... 16.71 At1g263... (applied)
At4g341... (not fou...
DRA_4965 15.71 At2g231... (applied)
DRA_3983 15.49 At3g232... (applied)
DRA_4259 13.85 At1g055... (applied)
5991GENE... 13.85 At1g055... (applied)
DRA_1681 12.67 At1g171... (applied)
5626GENE... 12.67 At1g171... (applied)
DRA_2019 12.26 At1g056... (applied)
Reaction (Reaction) [TP, AC]
Metabolite (Metabolite) [TP]
Enzyme (Enzyme) [AC]
TF (Transcription factor) [TF]
Protein (Protein) [TP, AC]
Gene (Gene) [TP, AC, DRA, TF]
EC (EnzymeClass) [AC]
Treatment (Treatment) [DRA]
ProtFam (Protein Family) [TP]
Protcmplx (Protein Complex) [AC]
Path (Pathway) [AC]
Thing (Thing) [GO]
PWM (Position Weight Matrix) [TF]
Comp (Compound) [TP, AC]
rg_by (regulated_by)
si_to (signals_to)
pr_by (preceded_by)
cat_c (catalyzing_class)
equ (equivalent)
ca_by (catalyzed_by)
in_by (inhibited_by)
is_a (is a)
m_isp (member is part)
p_isp (part is part)
cs_by (consumed_by)
pd_by (produced_by)
h_pwm (has_pwm)
co_by (cofactored_by)
en_by (encoded_by)
Console
Details
Legend
coordinates (347,1)
YOptions


ONDEX Visualisation ToolKit (OVTK) v0.2a
File Graph Layout Filter Options
Visualization
Enzyme
Envir onment
Genes
Path ways
Protein
Reac tion
Microarry Results
Layout Properties
Filter Properties
ID
expressionlevel
spotid
applied?
5627GENE... 180.2 At1g171... (applied)
DRA_1698 180.2 At1g171... (applied)
DRA_5103 61.42 At3g282... (applied)
4016GENE... 55.33 At2g294... (applied)
DRA_3821 28.67 At4g018... (applied)
DRA_1627 23.04 At2g154... (applied)
4286GENE... 23.04 At2g154... (applied)
DRA_1268 21.62 At2g413... (applied)
DRA_4240 17.65 At2g367... (applied)
3790GENE... 17.65 At2g367... (applied)
DRA_4084 16.87 At2g374... (applied)
DRA_1395 16.71 At1g263... (applied)
5390GENE... 16.71 At1g263... (applied)
DRA_4965 15.71 At2g231... (applied)
DRA_3983 15.49 At3g232... (applied)
DRA_4259 13.85 At1g055... (applied)
5991GENE... 13.85 At1g055... (applied)
DRA_1681 12.67 At1g171... (applied)
5626GENE... 12.67 At1g171... (applied)
DRA_2019 12.26 At1g056... (applied)
Protein (Protein) [TP, AC]
p_isp (part is part)
cs_by (consumed_by)
pd_by (produced_by)
h_pwm (has_pwm)
co_by (cofactored_by)
en_by (encoded_by)
Console
Details
Legend
coordinates (347,1)
YOptions
Integrated data from several databases
Filtering
Graph Analysis
Graph Layout


lignin biosynthesis
sinapaldehyde
sinapyl-alcohol
coumaryl-alcohol
coumaroyl-CoA
1.1.1.195
1.2.1.44
NADPH
NADP
coumaraldehyde
coniferyl aldehyde
Coenzyme A
coniferyl alcohol
feruloyl-CoA
CINNAMYL-ALCOHOL DEHYDROGENASE
CINNAMOYL-COA REDUCTASE
CINNAMYL ALCOHOL DEHYDROGENASE
At1g09500
At1g72680
At2g33590
cinnamyl alcohol dehydrogenase like (not available)
cinnamoyl-CoA reductase family (not available)
cinnamyl alcohol dehydrogenase putative (not available)
ABA (abscisic acid)
wound
sodium chloride
drought / dehydration / wilt

Associated to stress response genes




Regulated by transcription factors

# 3rd Integrative Bioinformatics workshop

4th to 6th September 2006

Rothamsted Research, Harpenden, UK

http://www.rothamsted.bbsrc.ac.uk/bab/conf/ibiof/

| | |
|---|---|
| 8th May 2006 | Paper submission deadline |
| 23rd June 2006 | Notification of acceptance for papers |
| 17th July 2006 | Camera ready paper submission deadline |
| 1st August 2006 | Registration deadline |
| 15th August 2006 | Poster submission deadline |