



D3.3 2nd version of the platform and Future work

Munich (October 2011)

Marc Poch, UPF (marc.pochriera@upf.edu)



Summary



- D3.3
 - Platform version 2 definition
 - New documentation
 - Web services
 - The Registry and myExperiment
 - Taverna and massive data
 - GRAF

[D3.3 > http://gilmore.upf.edu/WS/upload/files/414](http://gilmore.upf.edu/WS/upload/files/414)



Summary II



- Future work (preparing t30 workplan)
 - Web services
 - Taverna
 - The Registry and myExperiment
 - Massive data
 - New tools, new web services (your schemas)



Platform definition

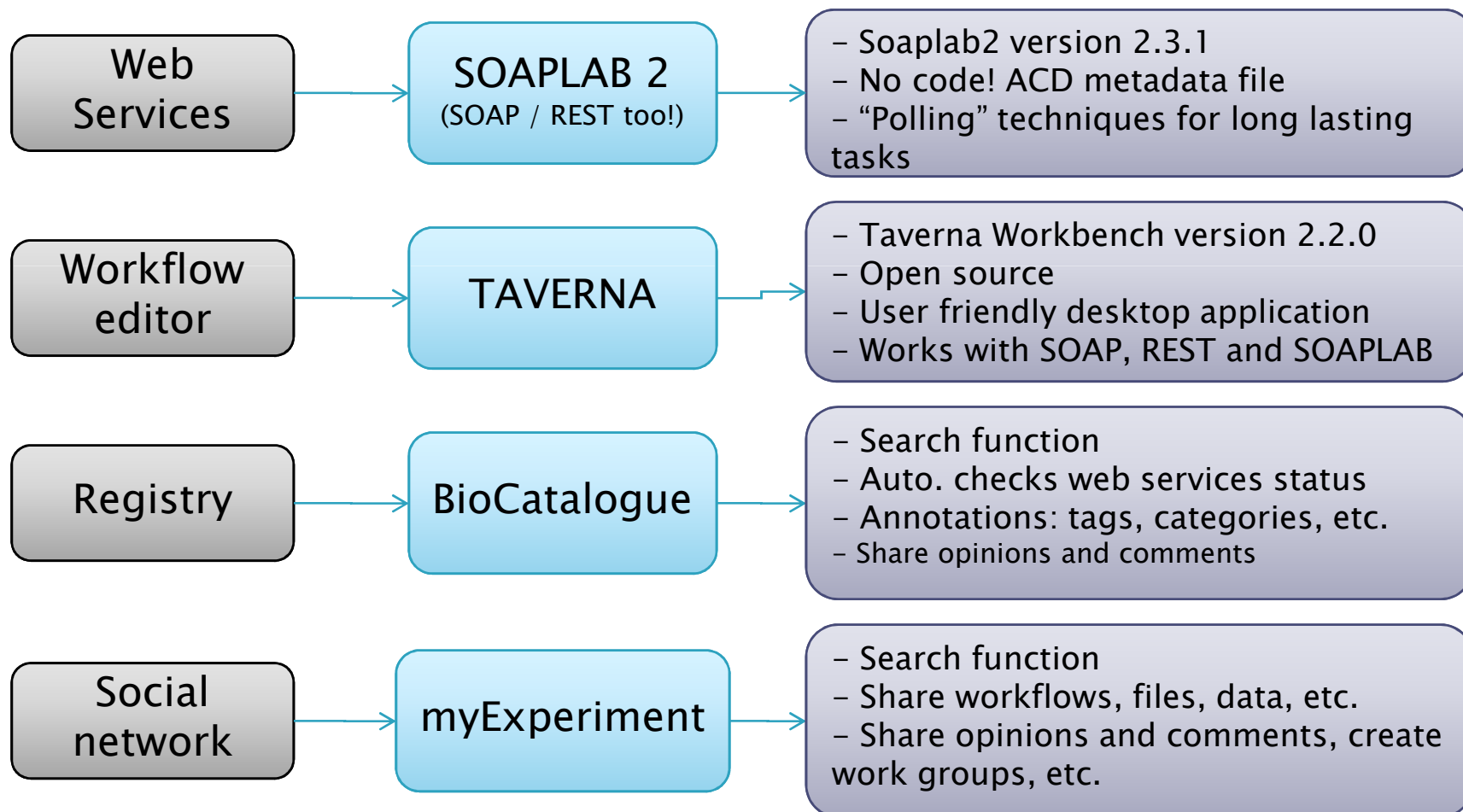


- The PANACEA platform is an **interoperability space** based on tools, guidelines, a Common Interface definition, and a “Travelling Object” specification
- Formal definition

Version 2:

- **Tools:** Taverna, BioCatalogue, myExperiment, Soaplab
 - **Common Interface:** WS interoperability
 - **Travelling Object:** TO1 (XCES) and GrAF
 - **Documentation**
- Technical Definition

PANACEA Platform version 2: already operational but not final.





New documentation



- TO1 schemas and Common Interfaces
 - <http://panacea-lr.eu/en/info-for-professionals/documents>
- GRAF only in D3.3
- PANACEA tutorial
 - <http://panacea-lr.eu/en/tutorials>
- myExperiment files
 - <http://myexperiment.elda.org/files>

Web services

- Soaplab
 - Tomcat 7 and Soaplab 2.3.1 bug (spinet bug)
 - Parameter name bug (found thanks to DCU aligners)
 - Soaplab output size limit patch (UPF)
 - <http://myexperiment.elda.org/files/3>
 - Soaplab polling
 - Limit web service usage (DCU)
 - <http://myexperiment.elda.org/files/4>
 - Temporary files mangement (ILC)
 - <http://myexperiment.elda.org/files/1>

- Soaplab wsdl validator

- http://ws04.iula.upf.edu/soaplab2-axis/#others.soaplab_wsdl_validator_row

CI report results:

There are only a couple of warnings reported by the validator script: 1) the **bilingual crawler** due to the fact that it's being validated against the CI of a monolingual crawler and obviously it fails for the two language parameters. 2) A **parser** web service which is only for Italian and the CI expects a language parameter that is not necessary in this case.

- Workflow

- <http://myexperiment.elda.org/workflows/25>



The registry and myExperiment

- <http://registry.elda.org/>
- New spinet link: directly to the web service
- Check status (how often?)

- <http://myexperiment.elda.org/>
 - New workflows and files

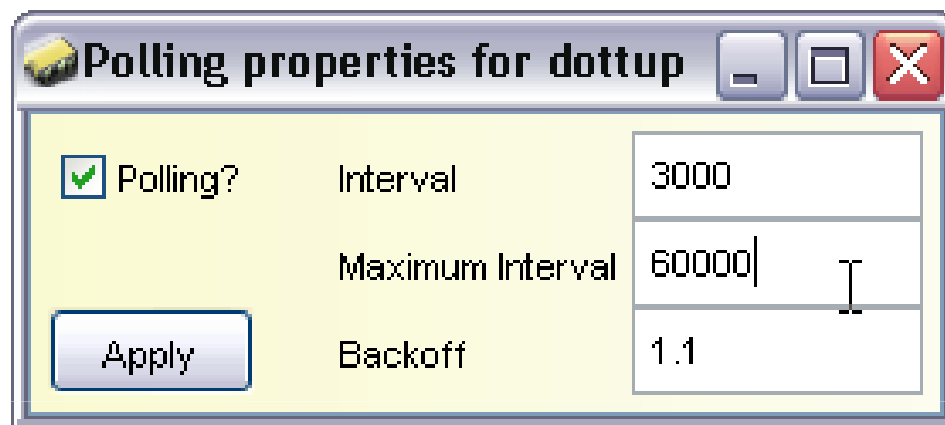
- Polling

- **Typical WS:**

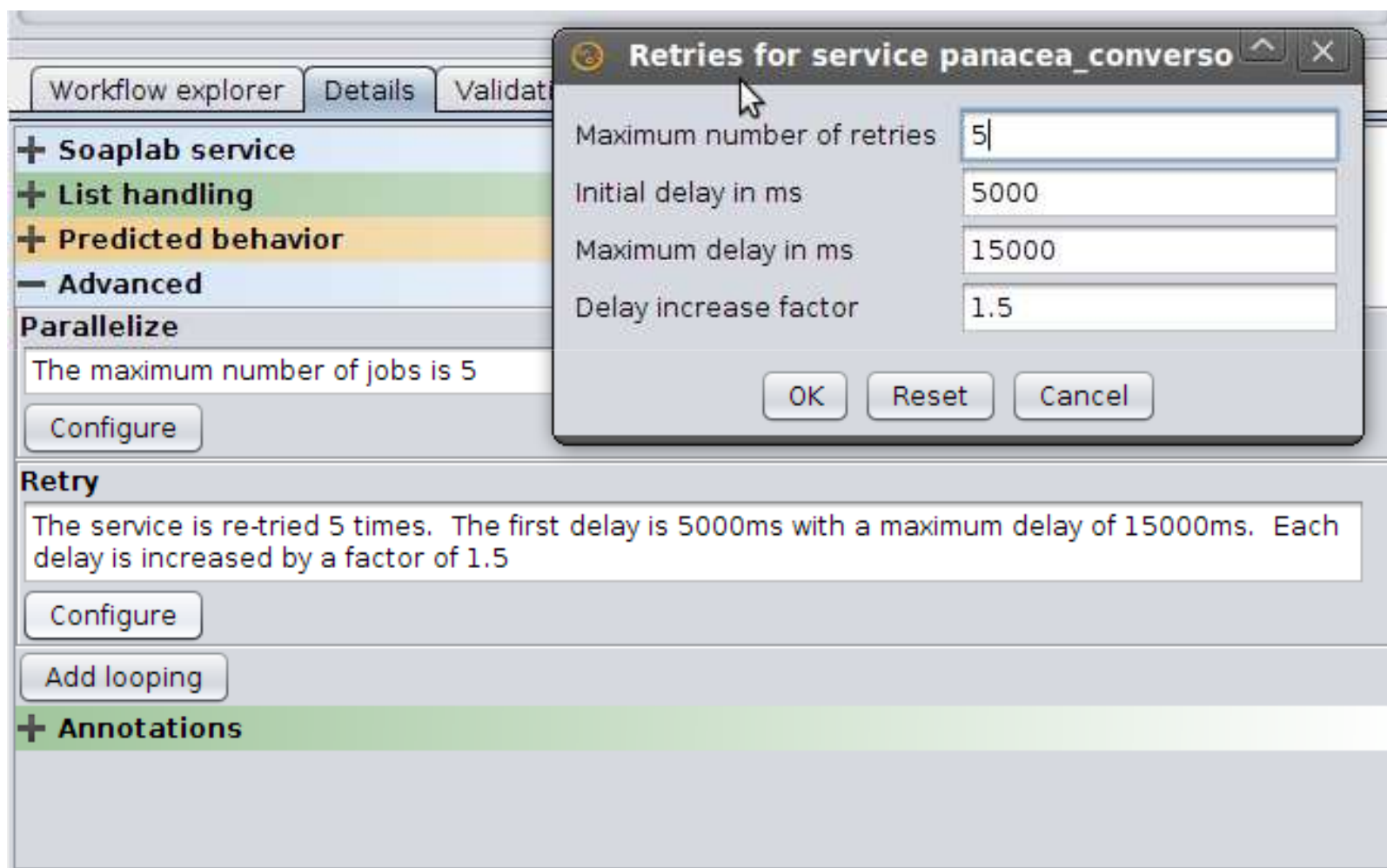
1. send request
2. wait. (if wait more than timeout it fails)

- **Polling WS:**

1. send request
2. Finished? (repeated every Interval with interval smaller than timeout)
3. ...
4. Get result



Retries



Workflow explorer Details Validation

- + Soaplab service
- + List handling
- + Predicted behavior
- Advanced

Parallelize

The maximum number of jobs is 5

Configure

Retry

The service is re-tried 5 times. The first delay is 5000ms with a maximum delay of 15000ms. Each delay is increased by a factor of 1.5

Configure

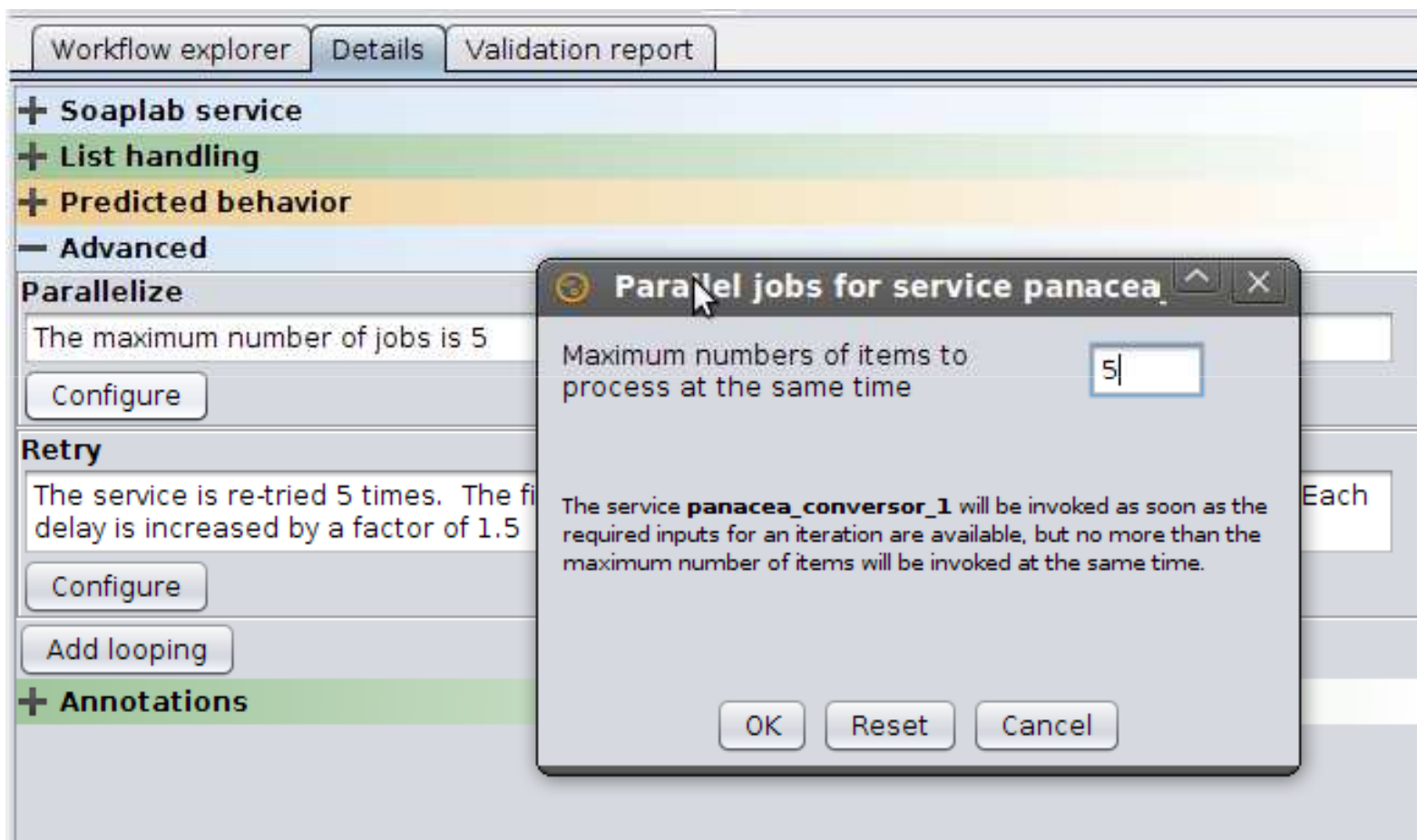
Add looping

+ Annotations

Retries for service panacea_converso

Maximum number of retries	5
Initial delay in ms	5000
Maximum delay in ms	15000
Delay increase factor	1.5

OK Reset Cancel



The screenshot shows the PANACEA interface with a configuration dialog box open. The dialog is titled "Parallel jobs for service panacea" and contains the following text:

Maximum numbers of items to process at the same time

The service **panacea_convertor_1** will be invoked as soon as the required inputs for an iteration are available, but no more than the maximum number of items will be invoked at the same time.

Buttons: OK, Reset, Cancel

The background interface shows a workflow explorer with tabs for "Workflow explorer", "Details", and "Validation report". The "Advanced" section is expanded, showing "Parallelize" (The maximum number of jobs is 5) and "Retry" (The service is re-tried 5 times. The delay is increased by a factor of 1.5).



Taverna Server



- Why a Taverna Server?
 - For long lasting workflows (you can turn off your pc)
 - More powerful server
 - Probably a better network
- Taverna Server 2.2.0
 - Not security, no graphical interface, few feedback (results, and provenance) etc.
- Wait for Taverna Server 2.3.0
 - Suposed to be released on August 2011



Taverna



- Encoding bug found by Linguatec:
 - Encoding problem related to read/write and parameters.
 - Reported to Taverna developers and fixed in Taverna 2.3.0 (not tested)
 - *Users can specify the character encoding and data type of input data (T2-1750)*
 - *“Read text file” can now take an encoding (T2-1750)*

Massive data

- Involved variables
 - Internet
 - The tools: deployed tools make diff. use of resources, memory leaks, etc..
 - The machine: memory, CPUs, HDs speed, etc.
 - Operating System
 - Web services
 - Tomcat: version, libraries used, parameters
 - Soaplab: version and parameters
 - Temporary files
 - Etc.



Handling massive data



- Taverna:
 - Re-design workflows: Retries, polling, parallelization, etc. Parameters optimization by empirical observation!
- Soaplab:
 - Use “Soaplab output size limit” and “limit web services”
- Temporary files:
 - Use scripts to clean your tmp files



Massive data report



- First report... (simple and to prove the use of retries, polling etc.)
- Second report: Stress tests on iula04v.upf.edu
- SCENARIO 1
 - VM, 4GB RAM, SAS HDs, Tomcat 6.0.29
 - A workflow with ONLY iula04v web services
 - 3 web services * parallelization x5 = 15 simultaneous processes
- ERRORS:
 - [Error bursts](#), [taverna hangs](#), etc.



Massive data report



- Scenario 2
 - **6GB RAM**
 - Taverna 2.2.0 and 2.3.0 (workbench and command line)
 - Taverna support team
 - Tomcat native libraries?



Massive data report



- Scenario 3:
 - **Tomcat 7.0.21 + libtcnative-1 1.1.20-1** (maverick repositories)
 - When it works... 5k docs > 2,5-4 hours
 - Found that linux directories have a 32k max folders
 - 5k docs * 3 WS = 15k folders!

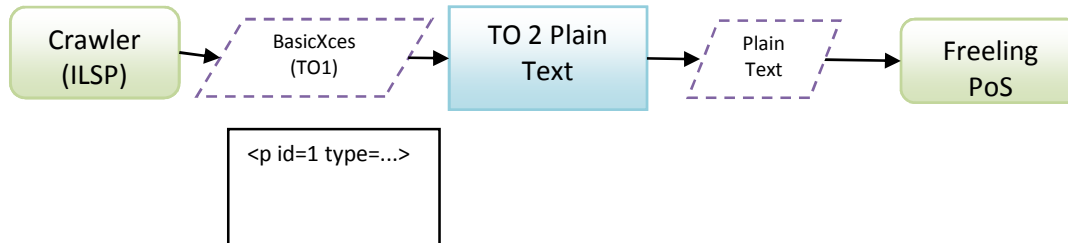
WEB SERVICE PROVIDERS:

- Updated software: **Tomcat + native libraries + monitor**
- Soaplab 2.3.1 with “**output size limit**”
- Temp files automatic management:
 - Check the **HD space**
 - Check the tomcat and soaplab **logs** (if errors, size can be GB!)
 - Check the **32k limit** in tmp folders
 - **Erase old** tmp and results files

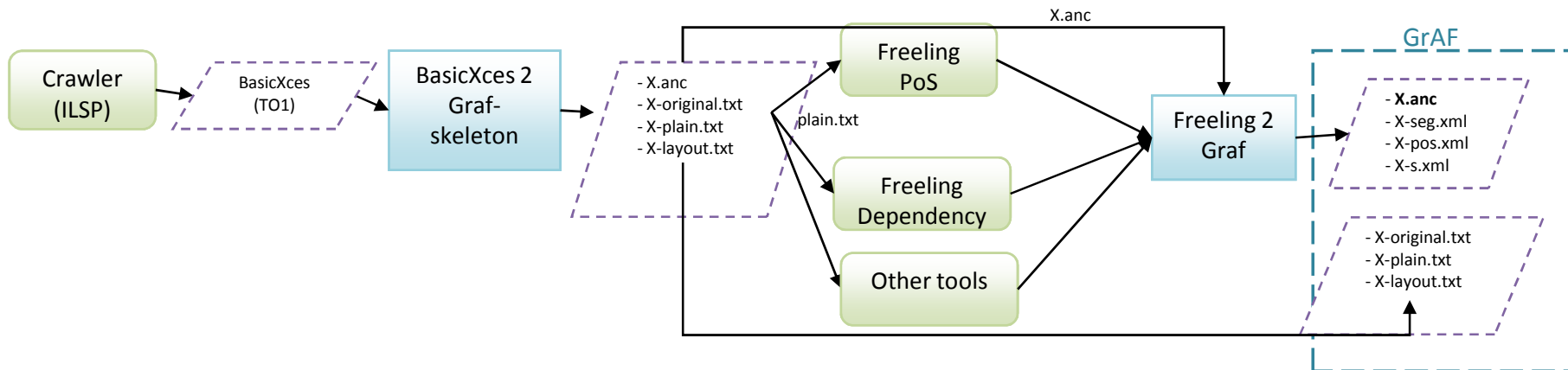
USERS:

- Taverna 2.2.0 works with Antonio! (maybe 2.3.0 too)
 - **In-memory** option activated! (faster and more robust)
- Workflows: **RETRIES! POLLING! And Parallelization** (careful!)

TO1 from crawler



GrAF: from BasicXces to GrAF





FUTURE WORK



- DoW: (t30): Third version (v3) of the integrated platform and documentation (v2 + WP4 PoS modules + WP5 Bilingual Dictionary Extractor + WP5 Transfer Grammar Extractor + WP6 Lexical Acquisition components + WF editor + Registry)
- **New tools** > new web services > new workflows > TO issues > converters > etc.
 - Your schemas will be used to predict issues and design workflows:
 - New web services?
 - New data formats?
 - Converters?
 - Corrections?



FUTURE WORK



The following topics will be integrated in the future work plan

- Deploy new ws: WP4, WP5 and WP6. (WP3 converters)
- Massive data (10k, 10M words, ok?):
 - Build new workflows (and update the old ones) following massive data recommendations:
 - Retries: 5
 - Parallelization: not more than 3 for common use (5 for stress test) per web service...
 - Not more than 8 processes in the same server (15 for stress) Depends on the tools!
 - Polling: for small files short intervals... some tests for optimal results.
 - Distribute crawled data (some): now everything is at ILSP
 - TESTS! All web service providers will design an stress workflow for their servers and do some tests
 - 3 web services minimum, etc.
 - <https://docs.google.com/spreadsheet/viewform?formkey=dHB6WWxpT2NuYmFvOFZZMW9yc1R0UHc6MQ>
 - Stress tester: (Probably ELDA, Olivier?)
 - Taverna workbench 2.2.0, Taverna workbench 2.3.0, Taverna command line 2.3.0
 - Stress workflows combining different web service providers

FUTURE WORK

- Taverna Server?
- Registry and myExperiment (finished?)
- GRAF:
 - UPF: modify freeing output dependency and chunk to work with stand-off
 - ILC: update grafconverter to use those outputs
 - Other partners will modify grafconverter if needed (deploy ws too)
- Web service providers:
 - **Tomcat with native libraries (libtcnative-1)**
 - **migrate to soaplab 2.3.2** + “**output size limit**” patch (UPF: 1k limit)
 - are DCU ws ok (naming and CI problems)? CI validation for everyone? is anyone already using soaplab 2.3.2?
 - **Temporary files**
 - Check the **HD space**
 - Check the tomcat and soaplab **logs** (if errors, size can be GB!)
 - Check the **32k limit** in tmp folders
 - **Erase old** tmp files
 - **Inotify?**



FUTURE WORK



- WP4, WP5, WP6 new ws and workflows
 - Are we forgetting something?
 - Data format? Converters problems?
 - Use your schemas to predict problems or missing things
- Are the Registry and myExperiment finished?
- Massive data:
 - Our goal is 10k docs (10M words aprox.) per experiment.
 - Is it OK?
- Any other ideas, problems, missing topics?



Thank you