



Project no. 018340

**Project acronym: EDIT**

**Project title: Toward the European Distributed Institute of Taxonomy**

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: “Global Change and Ecosystems”

## **C5.140 CDM-PESI integration accomplished and tested**

Due date of component: Month 49

Actual submission date: Month 50

Start date of project: 01/03/2006

Duration: 5 years

Organisation name of lead contractor for this component: 9 BGBM

Revision: Final

| <b>Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)</b> |   |   |
|--|---|---|
| <b>Dissemination Level</b>   |   |   |
| <b>PU</b>  | Public  | X |
| <b>PP</b>  | Restricted to other programme participants (including the Commission Services)        |   |
| <b>RE</b>  | Restricted to a group specified by the consortium (including the Commission Services) |   |
| <b>CO</b>  | Confidential, only for members of the consortium (including the Commission Services)  |   |

## CDM-PESI integration accomplished and tested

The Pan European Species directories Infrastructure (PESI, [www.eu-nomen.eu/psi/](http://www.eu-nomen.eu/psi/)) develops and integrated access system to the three major European taxonomic checklists Euro+Med PlantBase ([ww2.bgbm.org/EuroPlusMed/](http://ww2.bgbm.org/EuroPlusMed/)), Fauna Europaea ([www.faunaeur.org/](http://www.faunaeur.org/)), and the European Register of Marine Species (ERMS, [www.marbef.org/data/erms.php](http://www.marbef.org/data/erms.php)). Additional European checklists will follow as soon as the basic infrastructure has been established.

The backbone of the PESI infrastructure is an instance of an EDIT Common Data Model (CDM) Store, which is used to

- merge the participating checklists into a unified classification hierarchy,
- perform data quality procedures and report back to the checklist data providers,
- and create an export into the PESI data warehouse structure, which is a Microsoft SQL-Server 2008 database optimized for electronic publishing at both portal and web-service level.

In the context of this Component, we have implemented the entire information flow from the source checklists to the PESI data warehouse (see fig.). Import, merging, and export modules use and are integrated into existing CDM Java library belonging to the EDIT Platform for Cybertaxonomy.

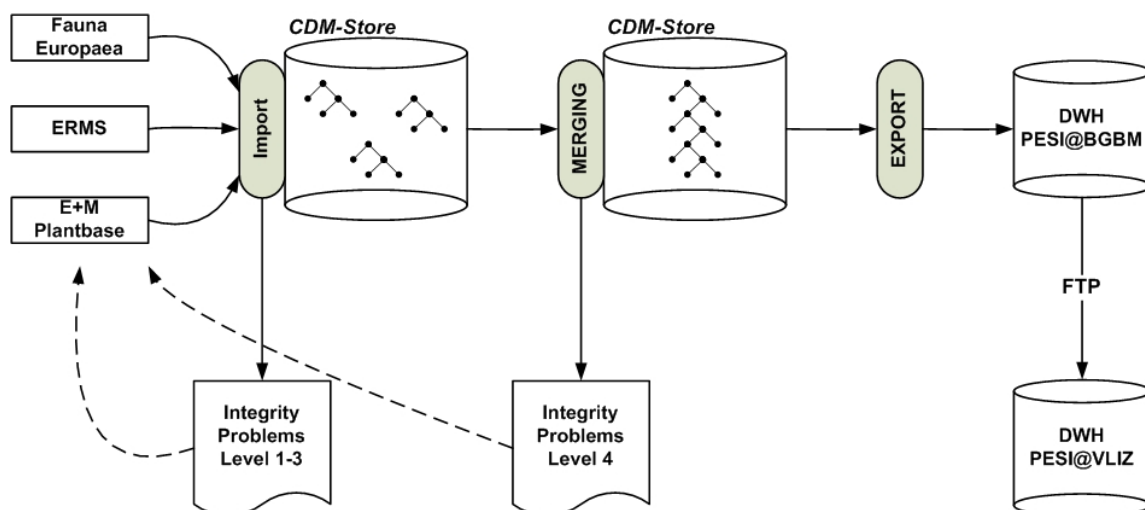


Fig.: PESI information flow from the contributing checklists to the data warehouse.

Data quality checks are implemented into the import procedure (for syntax, structure, and relations within the individual checklists) as well as the merging module (for integrity problems between checklists). The definition of data quality levels is following a PESI/EDIT specification compiled on a joint Wiki page (<http://dev.e-taxonomy.eu/trac/wiki/IntegrityRulesEditPESI>).

Followed by the initial phase of integrating the three major European checklists Fauna Europaea, ERMS, and Euro+Med PlantBase, further checklists will be considered for import into the e-infrastructure starting with Algae and Desmidiaceae. For this, we will define a standardized checklist data import format, which new data providers have to use as an export format. With this, we hope to be able to cover a variety of data import problems with only one additional import software module. A promising candidate for the intermediate and standardized data format is the Darwin Core Archive, which is also the agreed export format for PESI itself (<http://dev.e-taxonomy.eu/trac/wiki/IntegrityRulesEditPESI>).