FP7-KBBE-2010-4

Grant agreement no: FP7-265967

EuroFIR – Nexus

The EuroFIR Food Platform: Further integration, refinement and exploitation for its long-term self-sustainability

Deliverable 1.15

1st Report on Repository procedures and production

Due date of milestone or deliverable: M11

Actual submission date: M13

Start Date: April 01, 2011   Duration: 2 years

Organisation name of lead contractor for this deliverable or milestone: EuroFIR AISBL [IDUFIC]

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level (please check appropriate box)		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

EuroFIR Nexus WP1.3.5

Repository procedures and production

1ST Report prepared by Work Package 1.3.5

Ian Unwin (IDUFIC) and Susanne Westenbrink (RIVM)

30 March 2012

ABSTRACT:

Work undertaken on the EuroFIR Document and Data Repositories during the first year of the Nexus project is reported. The current framework of the Document Repository is described, with particular reference to the handling of conference posters. Initial proposals were prepared for a bespoke EuroFIR interface to the CiteXplore bibliographic database, but subsequently it was decided not to proceed with the development. Work to incorporate analytical reports is underway and aspects of Document Repository maintenance in collaboration with the EuroFIR Compiler Network are under discussion. Development plans for the Data Repository have been redefined following the decision to base the future system for its compilation on the FoodCASE food composition database management system.

Repository procedures and production
1st Report prepared by Work Package 1.3.5

Ian Unwin (IDUFIC) and Susanne Westenbrink (RIVM)

30 March 2012

Contents

Introduction

The concepts of the EuroFIR Document Repository and the EuroFIR Data Repository were developed within the five-year EuroFIR Network of Excellence (NoE) project. At the end of the NoE work, the development and operation of the two Repositories treated them as separate activities from the work of compilers to produce national compiled data, although the underlying motive was to reduce the amount of work that was duplicated by compilers in different countries. In the two years since the NoE work finished, the focus of further development has shifted to the facilities for compiling and retrieving data. As a result, during the first year of the follow-up EuroFIR Nexus project, it has been decided to concentrate on developing the major systems that compilers will use for the compilation (FoodCASE) and searching (eSearch) of food composition data. Production of, and access to, the Repositories will as far as possible be integrated into the FoodCASE and eSearch applications.

This report documents the work on the Repositories during the first year of the EuroFIR Nexus project and outlines the implications for further development of integrating their operation with the use of the FoodCASE and eSearch programs.

EuroFIR Document Repository

Overview and framework

Figure 1: EuroFIR Document Repository framework
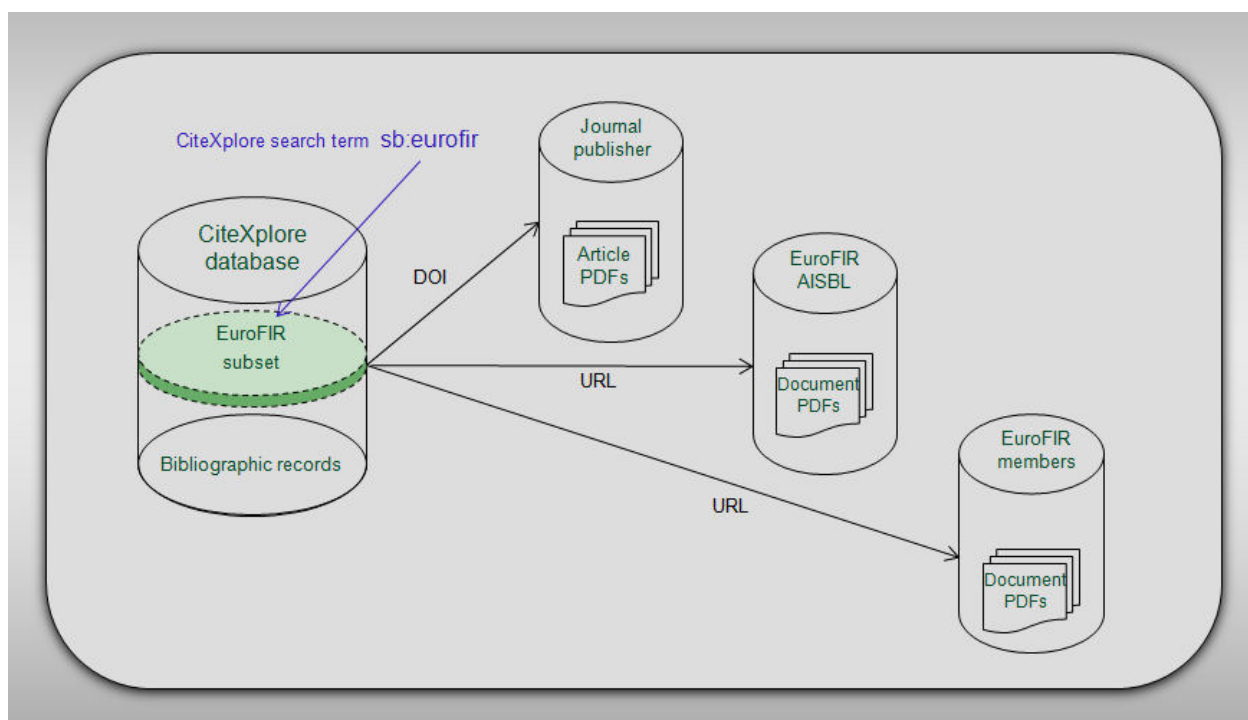
EuroFIR is building a Document Repository of bibliographic references relating to food composition information (Unwin *et al.*, 2009; EuroFIR, 2010A). As shown in Figure 1, the Document Repository framework consists of two elements:

- bibliographic records held in EBI's CiteXplore bibliographic database[1], flagged as belonging to the Document Repository, and linking to:
- the original documents if available online, usually as PDF files.

---

[1] http://www.ebi.ac.uk/citexplore/

EBI's CiteXplore database is a collection of bibliographic records for over 25 million documents, including the full PubMed database, together with Agricola records for selected food-related journal titles and records added by EuroFIR for documents relevant to food composition work. Any CiteXplore record may be assigned a special indexing term that indicates it belongs to the EuroFIR Document Repository, as being relevant to food composition work.

Figure 2: CiteXplore record for poster

Figure 2 shows the information held in a CiteXplore record, as input under the EuroFIR project. This poster record was retrieved using the query shown in the search box, i.e. "sb:eurofir src:ctx pub_type:poster norwegian". The query term "sb:eurofir" retrieves records from the EuroFIR subset of CiteXplore. "src:ctx" selects records input directly to CiteXplore, although this may be a redundant term in this case as most posters, retrieved by ""pub_type:poster", will have been added directly. The term "norwegian", without a prefix, is treated as a general query term and locates the word in a range of fields, including the title and abstract as shown by the green highlighting in Figure 2.

This bibliographic record includes an abstract and relates to a poster held on the website of a member of the EuroFIR Compiler Network. The CiteXplore records contain links to full-text versions of the documents when these are accessible over the World Wide Web. In this case the PDF file of the poster is held in the *matportalen.no* portal of the Norwegian health and food authorities. The green border to the link icon indicates that the document is freely available. If the icon has a red border, the organisation holding the document file has placed restrictions on access to the document, for example subscription based access to journal articles.

Content status

At the end of February 2012, the total number of CiteXplore records flagged as belonging to the EuroFIR Document Repository was 11162. Table 1 lists the content by the type of document and indicates the effect of flagging all documents for the two journals *J. Food Compos. Anal.* and *Food Chem*.

Table 1: Document Repository content

Document type and source		Number
	J. Food Compos. Anal.	1366
Journal article	Food Chem.	9214
	Other journals	275
Book		113
Book article		15
Report		122
Report article		2
Poster		55
Total		**11162**

Nexus year 1 activity

Journal scanning

In addition to the full coverage of journal articles published in *J. Food Compos. Anal.* and *Food Chem.*, nine further core journals were scanned for relevant articles under the earlier EuroFIR NoE project. As described in EuroFIR (2010A), scanning was completed for journal issues published in 2009.

At the beginning of the Nexus work, plans were discussed to undertake journal scanning for issues published over the period 2010 to 2012. Further, it was proposed to launch the Document Repository during the second half of 2011, once the scanning of the 2010 material had been completed.

However, after further consideration of the resources required and the benefit of the work, it was decided that the effort involved was not justified by the potential results. Routine scanning of a small number of journal titles did not produce a coherent set of relevant articles representative of the overall literature. Searching of this would be adequately served by querying the full CiteXplore database. Since this holds records sourced from PubMed and Agricola, as well as further records relevant to food composition added by EuroFIR, its coverage of food composition material is much superior to other openly available bibliographic databases.

On the other hand, having a EuroFIR subset of food composition articles flagged in CiteXplore does enable EuroFIR users to target the food composition literature more effectively. A practical strategy for building the Document Repository collection of relevant journal articles is to be developed in the context of the EuroFIR Compiler Network. This may involve procedures by which any compiler in the Network can submit suitable journal article citations to be flagged in CiteXplore. Although this may include any material that the compiler considers of interest, at a minimum it should require that all source references for values recorded in a compiler's Food Composition Database (FCDB) are included in the Document Repository. This more restricted approach may avoid one potential problem with the submission of articles as they are published, whereby several compilers spend time submitting the same articles at the time that it is published. Alternatively, individual compiler might be assigned journal titles that they normally

read anyway, so that only they submit articles as they are published. This would amount to a more flexible system than the previous procedure, which defined the work as separate from normal compiler activities. Redefining it in this way may also make the work more easily sustained by EuroFIR after the Nexus project finishes.

CiteXplore input for member additions

The Czech Institute of Agricultural Economics and Information requested the addition of 12 documents to the Document Repository. The Institute provided bibliographic data in the appropriate template specified in EuroFIR (2010A) and IDUFIC used this to create corresponding CiteXplore records using the EBI web-based submission system. EBI were sent the list of new record identifiers to be flagged as part of the EuroFIR subset.

The EuroFIR Compiler Network is planned to become active during April 2012. It is suggested that an early activity should be an offer to include documents submitted by members in the Document Repository. As well as the submission of relevant journal articles, as discussed in the above section, the documents would include internal and national reports that are normally not as easily accessed worldwide as the journal literature.

PosterBoard production

The EuroFIR PosterBoard service has been developed to improve the availability of posters presented at conferences or elsewhere. Hitherto, such posters have been available to be viewed at the time that they are presented, but have been difficult to retrieve and view later as part of the overall dissemination of scientific information. Some aspects of the wider dissemination of conference posts were noted in presentations prepared in June and September 2011 (IDUFIC/EuroFIR, 2011; Unwin *et al.*, 2011).

Posters presented at the 3rd International EuroFIR Congress held in Vienna in September 2009 were used to prototype PosterBoard production. PDF files for 77 of the posters were made available by their presenters and are located with the Document Repository area of the EuroFIR server. The file names were standardised to consist of the poster's conference reference and the first author surname. A Congress posters page was added with the Document Repository area of the EuroFIR website and the available posters listed on it. The list provides links to the PDF and to the corresponding CiteXplore record, which was created using the EBI submission tool. The linked CiteXplore record allows the full bibliographic data to be viewed, so the information on the webpage is limited to the poster reference and its title (with the first author name viewable by positioning the mouse over the PDF link).

The CiteXplore record includes a Publication Type indicator of *Poster* (as shown in Figure 2). This allows a search query to specify that records for posters are retrieved (or alternatively they can be excluded from the results). The provisional decision was taken not to include abstracts in the CiteXplore record. This would need considerable resources to write an accurate short abstract and long abstracts such as those provided in conference proceedings might over-weight the document in retrieval in comparison with the journal literature and other sources.

Posters presented at the 9th International Food Data Conference (9IFDC) held in Norwich in September 2011 have also been requested from their presenters. So far, 45 posters have been received and as of March 2012 work is progressing to add these to the Document Repository, using the procedures outlined above for handling the Vienna posters.

CiteXplore record archiving

Records added to the CiteXplore database should be archived by EuroFIR so that they can be loaded to an alternative bibliographic database, for example if the CiteXplore system should at any time become unavailable to EuroFIR. Optimal procedures need to be developed for this, but a provisional procedure has been operated. The 505 records that have been input through the EBI submission tool and are flagged as in the EuroFIR Document Repository have been retrieved using the CiteXplore query "sb:eurofir src:ctx" and exported in two formats, those for

the Endnote and Reference Manager applications. The archive files have been stored within the Document Repository area of the EuroFIR server, in a sub-directory named *CTXarchive*.

Development of Document Repository

EuroFIR interface to CiteXplore

The final report on the Document Repository from the EuroFIR NoE project (EuroFIR, 2010A) proposed the development of a bespoke interface for searching the CiteXplore database, allowing searches to target the EuroFIR subset of records. This would be specifically available to EuroFIR members and subscribers. Such an interface would also allow the implementation of facilities targeted at food composition and compiler requirements, for example requesting of journal articles and other documents to be added to the Document Repository or providing details of compositional values reported in a reference, such as a national food composition database.

An initial description of the requirements for the CiteXplore interface was prepared (Unwin, 2011). However, review of the development resources available within the Nexus project and prioritisation of planned tasks concluded that development of the interface would not be undertaken at this stage.

Analytical reports

It has been decided that the Document Repository work should give priority to making available to the compiler community the results of food analysis projects reported in analytical reports that are currently not easily accessible and thus are considered *grey literature*. Initial work will be directed at adding analytical reports cited in the Dutch NEVO and UK food composition databases to the Document Repository. The data in these reports will also be considered as possible pilot data to be used in the development of the Data Repository.

RIVM have prepared bibliographic data for 9 reports cited in the NEVO database. The reports are currently in various forms, for example as Excel spreadsheets, and work is ongoing to decide how these should best be presented as accessible documents. In some cases, permission from the producers of the reports must be sought for the documents to be made available through the World Wide Web. It is proposed to hold the original documents in the Document Repository area of the EuroFIR server. Once the documents are accessible, records for them will be added to CiteXplore.

Developing bibliographic expertise

Preparation of bibliographic data for Document Repository items and their input into CiteXplore require a substantial level of knowledge in the selection, editing and formatting of the information to be included. Most working scientists have the skills to search the literature and prepare publications, possibly supplemented by help from support facilities such as information and library services. However, the preparation of high quality bibliographic data requires detailed knowledge of conventions, policies and data structures much beyond that necessary in general scientific work. Even if the expertise is available within a compiler's organisation, this may not be accessible for supporting EuroFIR Document Repository production.

The best option, if it is available, would be to involve one or more information and library departments of Compiler Network organisations in the Document Repository work. Investigating possibilities for this might be considered a high priority once the Compiler Network becomes fully operational. One future requirement for the work will be the introduction of extra CiteXplore record curators and EBI may prefer that these do have the well-based bibliographic experience of information specialists.

Another aspect of building and maintaining the knowledge necessary for creating a quality Document Repository is clear and detailed documentation of the policies and procedures involved. As well as general policies and procedures, it should specify the information content

and format of the fields within the bibliographic data of the various document types. The options for maintaining effective documentation, e.g. the use of a Wiki, should be investigated. It is preferable that the documentation is available to all users of the Document Repository as this will deepen the knowledge of the information being searched and viewed. It has the added benefit that informed users might progress to become effective contributors to the Document Repository work.
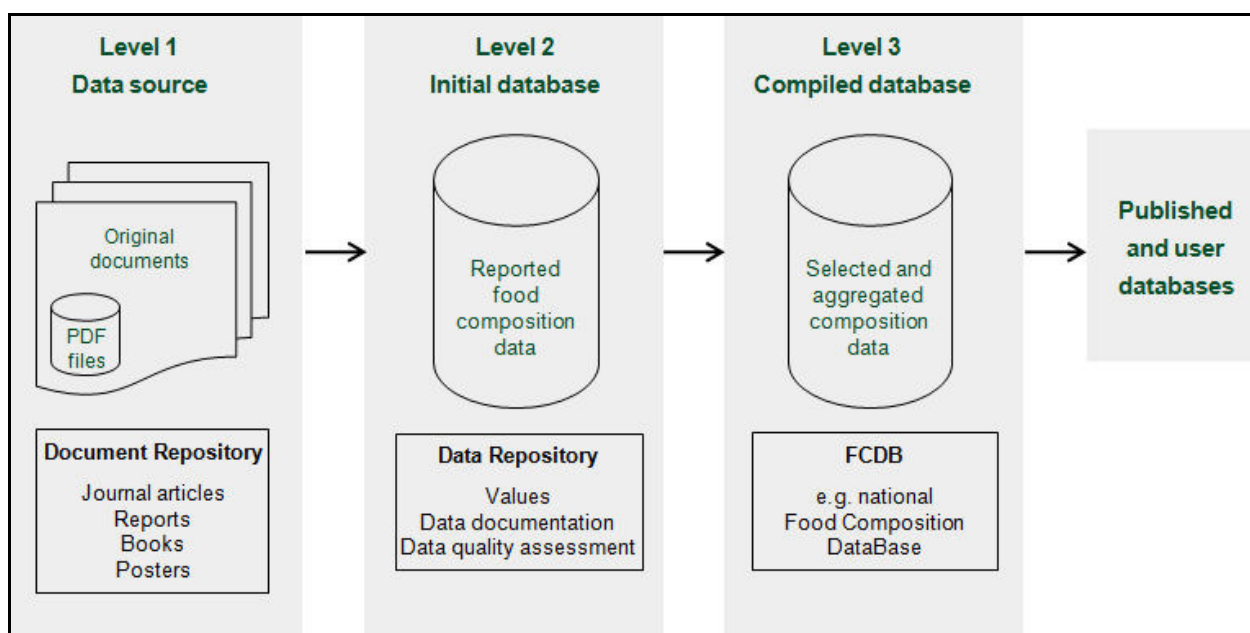
EuroFIR Data Repository

Overview and framework

The concept of the EuroFIR Data Repository was described in a deliverable of the EuroFIR NoE project, together with options for its development and implementation (EuroFIR, 2010B). Since those proposals, ideas have evolved considerably, as outlined in Unwin *et al.* (2011) and below.

The concepts of the Document and Data Repositories were developed in response to practical considerations, in particular to reduce the amount of work on reported data duplicated by the various data compilers. However, they also match well with the concept of data levels in the processing of food composition data introduced by Greenfield & Southgate (2003, pp. 10-12) and further developed in Becker *et al.*, (2007, pp. 4-5). Figure 3 outlines the relationship of Document and Data Repositories to the data levels framework and the overall process of food composition data compilation.

Figure 3: Food composition data levels and the EuroFIR Repositories

Data levels 1 and 2 correspond to the Document Repository and the Data Repository, respectively. Each of these two steps in the overall data compilation process involves work on the available literature, collecting relevant documents and extracting the food composition information reported. Until now, each compiler of a national FCDB has undertaken these tasks independently, thus duplicating the effort involved in the first two steps of the compilation work. The two Repositories allow the work to be performed just once, with the results available to be shared by other groups belonging to the Compiler Network.

The preparation of level 2 data for the Data Repository consists of three aspects, capture of the reported values, documentation of the values and quality assessment of the values. Possible approaches to the capture of values using optical character recognition procedures, thus avoiding data re-keying, were considered in EuroFIR (2010B). Data are to be documented in

accord with the EuroFIR standard procedures (Becker *et al.*, 2007; Becker *et al.*, 2008). Guidelines for quality assessment and the assignment of a Quality Index for data are provided in Oseredczuk & Salvini (2008). The resulting documented and assessed data will be available for searching, viewing and importing into the food data management systems (FDBMS) of compilers as values contributing to the compiled data in the level 3 FCDBs.

Development proposals

Development work on the Data Repository under the EuroFIR NoE project concentrated on the initial, data capture step of data preparation (EuroFIR, 2010B). It was proposed that data files would be prepared in spreadsheet format in a standard tabular layout suitable for importing into an FDBMS. The components would be indicated by using the EuroFIR component identifiers as column headings and the foods would be identified by their names and LanguaL indexing. Each individual data file might be marketed as a separate item and thus initial work was planned in terms of the preparation of a given number of prototype files. Adding data documentation and quality assessment were seen as later developments, once production of the basic data files had become operational.

During the two years since these initial plans were proposed, two key aspects relating to the management and dissemination of food composition data have advanced towards a collaborative environment. A web-based FDBMS, FoodCASE, has been introduced as a EuroFIR software tool, the use of which is being prototyped by several compilers (Colombani & Presser, 2010). Secondly, the prototype eSearch tool, which allows users to search simultaneously all participating FCDBs, is being enhanced within the EuroFIR Nexus project. Thus the applications for both compiling and retrieving food composition data, central to the work of compilers, have become major foci for EuroFIR development effort. Since the building and searching of the Data Repository will be best performed using these software tools, it is logical to develop the Data Repository procedures within these environments. A further advantage is that the systems and procedures for managing the Repository data will be familiar to compilers using the same systems for their FCDB work.

FoodCASE has recently become available for the development and testing of Data Repository compilation procedures. Key aspects are the options for data import, for example for data values held in a spreadsheet format, and the ability to export files in the EuroFIR Food Data Transport Package format for searching by eSearch. Documentation of the values would be undertaken in FoodCASE, but at present this may need to be done on a value-by-value basis. Procedures for the bulk documentation, for example of analytical method, may be required for data from a single document. The method used for each component will generally be the same and thus a facility allowing the method to be associated with a component would greatly improve the efficiency of data documentation. Such a facility would be equally useful in FCDB compilation and thus developing FoodCASE to support extra options for documenting data would benefit all its users. The quality assessment of Data Repository data will initially be given lower priority in development and implementation, but can be introduced as and when it comes into general operation as part of preparing level 3 (compiled) data.

Analytical reports are the basic source of food composition data and are seen as being of high priority for inclusion in the Data Repository. It is planned to use reports associated with the national food composition efforts in the Netherlands and the UK as prototype material. In some cases, level 2 (reported) data may already exist in national FoodCASE implementations and this will be investigated further, initially in the Dutch NEVO database. The initial objective is to create a prototype Data Repository database of documented data that will make useful data available as well as to develop the compilation procedures and demonstrate the retrieval of reported food composition data. A small set of test data should be made available as soon as possible so that any implications for the enhanced eSearch specification can be identified.

Concluding remarks

At the start of the EuroFIR Nexus project, some work was done in the furtherance of the earlier approach of independent development of the Repository activities and services. An initial specification was prepared of the bespoke CiteXplore interface for the Document Repository and plans projected the production of individual data files as the initial activity of the Data Repository. The decisions to revise these plans were only ratified well into the first year of the Nexus project, but it has been possible to assimilate the new approach into the overall development and operation of the Repositories. This integration of the work with the software environments routinely used by compilers significantly increases the viability of the work from a variety of operational aspects, as well as drawing benefit from the major software developments that are underway. Repository work can be seen as a collaborative effort that avoids duplication of effort by compilers without requiring significant extra resources. Such collaboration is central to the continuing success of the EuroFIR AISBL and its Compiler Network.

References

Becker W., Unwin, I.D., Ireland, J. and Møller, A. (2007). *EuroFIR Draft Standard Description.* Proposal for a EuroFIR standard on food composition data, version dated 13 July 2007, EuroFIR Project Management Office, Norwich, UK. Available at: http://www.eurofir.eu/sites/default/files/docrep/guidelines/EuroFIRStandardDescription(2007-07-13).pdf.

Becker W., Møller, A., Ireland, J., Roe, M., Unwin, I.D. and Pakkala, H. (2008). *Proposal for structure and detail of a EuroFIR Standard on food composition data. II. Technical Annex, Version 2008.* EuroFIR Technical Report D1.8.19, EuroFIR Project Management Office, Norwich, UK. Available at: http://www.eurofir.eu/sites/default/files/docrep/guidelines/EuroFIRStandardAnnex(2008-12-31&D1-8-19).pdf.

Colombani, P. and Presser, K. (2010). *Usability testing of the single value module and recipe module of FoodCASE and recommendations for future use by EuroFIR AISBL.* EuroFIR Technical Report D1.8.41, EuroFIR Project Management Office, Norwich, UK. Available at: http://www.eurofir.eu/sites/default/files/docrep/tools/FoodCASETesting(2010-04-19&D1-8-41).pdf.

Greenfield, H. and Southgate, D.A.T. (2003). *Food Composition Data: Production, Management and Use.* 2nd Ed., FAO, Rome. Available at http://www.fao.org/infoods/publications_en.stm.

EuroFIR (2010A). *Report on EuroFIR Document Repository status and future compilation.* EuroFIR Deliverable D1.8.34; 9 June 2010.

EuroFIR (2010B). *EuroFIR Data Repository options and development proposals.* EuroFIR Deliverable D1.8.35; 10 May 2010.

IDUFIC/EuroFIR (2011). *EuroFIR PosterBoard.* Slides prepared June 2011. Available at: http://www.eurofir.eu/sites/default/files/docrep/background/EuroFIRPosterBoard(2011-06).pdf

Oseredczuk, M. and Salvini, S. (2008). *Guidelines for quality index attribution to original data from scientific literature or reports for EuroFIR data interchange.* Report prepared by EuroFIR Work Package 1.3, Task Group 4. Available at: http://www.eurofir.eu/sites/default/files/docrep/guidelines/EuroFIRQIguidelines(2008-06-30).pdf.

Unwin, I.D. (2011). *EuroFIR interface to CiteXplore – initial considerations.* Report dated 7 June 2010. Available at: http://www.eurofir.eu/sites/default/files/docrep/background/ EuroFIR-CTXinterface(2011-06-07).pdf

Unwin, I.D., Bell, S. and Finglas, P.M. (2011). *Data and Document Repository.* Presentation to the EuroFIR Nexus 1st Annual Meeting, Norwich, UK, 12th & 13th September 2011. Available at: http://www.eurofir.eu/sites/default/files/docrep/background/EuroFIRDoc&DataRepos(2011-09-11).pdf

Unwin, I.D., Colombani, P., Moller, A., Roe, M.A., Stoehr, P.J., Westenbrink, S. and Finglas, P.M. (2009). *EuroFIR Document Repository: a new approach to accessing the food composition literature*. Poster presented at the 3th EuroFIR Congress, 8-10 September 2009, Vienna, Austria. Available at:
http://www.eurofir.eu/sites/default/files/docrep/poster/eurofir/2009congress/1_34_Unwin.pdf