diXa data warehouse: Exploring data from toxicogenomics studies

Vera Matser [1]

- Gene Expression
- Ontologies
- Chemical biology
- Beginner
- 1 hour

The diXa data warehouse is a curated collection of toxicogenomics experiments, accessible through the diXa website [2]. This course will show you how to use the diXa warehouse and what data is available.

Learning objectives:

- Learn what the diXa data warehouse is and where the data comes from
- Be able to access and navigate the diXa data warehouse
- Search the diXa data warehouse for a chemical compound or a study

What is toxicogenomics?

Understanding the molecular mechanisms of toxicity

Toxicogenomics (1,2 [3]) is the study of the structure and function of the genome as it responds to cellular exposure to foreign agents. It combines the field of toxicology with high-throughput technologies such as transcriptomics [4], proteomics [5] and metabolomics [6] in an attempt to understand the molecular mechanisms of toxicity. Studying these mechanisms can help in building better animal and human models, with the goal of better predicting toxic response in humans.

Moving away from animal models towards in vitro assays

Toxicology has become a data-dense discipline with the opportunity to use high-throughput genomic technologies. Under initiatives such as the NC3Rs, the aim is to refine and reduce animal-based testing models with the use of non-animal in vitro/in silico assays, ultimately with the goal of replacing animal use. These assays/models developed as a result of toxicogenomics could be better predictors of human toxicity in vivo, less costly and socially
more acceptable.

What is the diXa data warehouse?

The diXa Project

The Data Infrastructure for Chemical Safety (diXa) project [7] is an EU FP7 funded initiative to develop a single resource to capture data produced by toxicogenomics [8] studies, and to ensure sustainability of this resource for use by the wider research community.

What is the diXa data warehouse?

The diXa data warehouse is a curated collection of toxicogenomic experiments with cross-links to external resources including chemical/toxicological and human disease databases [3].

The aim of the warehouse is to provide:

- Accurate representation of chemical compounds;
- Easy discovery of studies of interest;
- Facilitation of data analysis.

The warehouse will store (or link to) the raw and processed data [9], and to their corresponding metadata [10] (e.g. sample characteristics, technologies used, type of measurements, etc).

Why do we need the diXa warehouse?

The diXa warehouse is important because it will:

- Provide a central repository for toxicogenomics data;
- Provide large and well curated experimental datasets for scientists to build computational models;
- Safeguard the data produced by EU toxicogenomics projects for future use.

When to use the diXa warehouse

You can use the diXa data warehouse:

- As a central repository of toxicogenomics studies;
- As a starting point to investigate compounds used and analysis performed in toxicogenomics studies.

You cannot use the diXa data warehouse:

- To provide an analysis pipeline for your data;
- As a general repository of functional genomics [11] experiments. If this is what you are looking for then ArrayExpress [12] may be more relevant to you.
How to access the diXa warehouse

The data warehouse can be accessed through the diXa project website [13]. You do not need any login credentials and all information is freely available (Figure 1).

![Figure 1](image1.png)

**Figure 1** Accessing the diXa data warehouse from the diXa project page.

Navigating and searching the diXa warehouse

There are two main ways you can search the diXa warehouse (Figure 2):

- Using the search bar at the top of the page;
- Using the browse icons in the centre of the page.
The browse icons provide access to data on studies, compounds, analysis or disease. The number in the right bottom corner of each icon represents the total number of entries for each category.

Figure 2 The diXa warehouse homepage.

Searching

If you want to search the diXa warehouse using a keyword, you will need to select either the studies, compounds, analysis or disease data tab in the top right of the page before entering your search term (Figure 3).

The search function is ontology-enabled and uses the Experimental Factor Ontology (EFO) developed by EMBL-EBI.
Figure 3 Searching the diXa warehouse.

**Browsing**

If you don’t have a specific search term in mind you can browse the available datasets by selecting one of the main tabs (without adding a keyword). An example is shown below in Figure 4.
Participating projects

An additional method of exploring data within the warehouse is to search by participating project. These are listed in a table at the bottom of the homepage, giving an overview of all EU and other projects that have contributed to the diXa warehouse (Figure 5).
Figure 5 Participating projects table.

Mouse-overs will give you more information on the project. If you click on the project name it will take you to the studies available for that project.

Studies, analyses and disease data

When you browse or search for studies, analyses or disease data you will be presented with a list of results, similar to the one shown in Figure 6.
Figure 6 The study view list.

You can find out more about the study, analysis or disease data you are interested in by clicking on the relevant ID link. You can also sort the columns by clicking on the blue headers at the top of the list.

Studies

The study details page provides all the information that has been stored in the diXa warehouse for the selected study, including the names of the chemicals used. There are also links to the data files and original EMBL-EBI repositories such as ArrayExpress [14] (Figure 7).
Figure 7 An example of the study details page.

The ontological terms used in the study appear as clickable links, allowing you to search the warehouse for any of these terms - e.g. all studies performed on the Agilent technology platform.

If further analysis of this study is available in the diXa warehouse, you will find a link at the bottom of the page (as shown in Figure 7).

Analysis

To ensure that the information contained in the diXa warehouse is compatible, complete and accurate, robust quality control (QC) and validation steps are carried out on the data and metadata [10]. The results of QC data are available through the analysis section of the diXa warehouse.
Depending on the project, the analysis can cover one or multiple studies. The studies covered in the analysis are listed at the bottom of the analysis details page (Figure 8).

![Analysis Details](image)

**Figure 8** The analysis details page.

**QC files**

All the QC files for a given dataset are available on the analysis details page (Figure 8). However, you can get a quick overview of the dataset by looking at the summary, which describes the assay [15], the quality assessment, and the observed effects of batches, donors, time, and concentration.

The data overview (PDF file) provides you with additional information and figures for the distributions of the gene expression [16] intensities, box plots, and the performed principle component analysis [17] (PCA).

**Disease data**
The toxicogenomics [8] studies in the diXa warehouse are complemented by publically available disease datasets. The 188 disease datasets fall into three sub-categories: liver, kidney and cardiovascular disease.

The majority of the available disease data are mRNA expression profiles [18], but some other data types are available (e.g., miRNA [19] expression data, single nucleotide polymorphisms/genome wide association studies, mutation studies, copy number variation studies, and DNA methylation [20] data).

Searching for disease data

The diXa warehouse includes ICD-10 [21] terms for the human disease datasets. You can search the disease data by organ, by using a disease-related term or by entering the numeric or the textual part of an ICD-10 term.

Figure 9 Disease details page.

The disease details page (Figure 9) provides you with the source of the data, a short study description and the ICD-10 terms (which can be used to start a new search query). If applicable, the links to the original datasets in ArrayExpress [14] are provided, together with the re-annotated metadata [10] files that include ICD-10 terms and codes.
The Comparative Toxicogenomics Database (CTD) provides information on whether a specific compound is likely to cause/interfere with a particular disease. The compounds represented in the diXa warehouse were screened for a relation to the liver, kidney and/or cardiovascular system using the CTD database. If an association was present for one of these ‘organs’, you will find a link to the human disease data for that particular organ on the compound details page (Figure 11 [22]).

Exploring chemical compounds

Compound list view

You can search for a compound by entering the compound name, a valid synonym [23], standard InChI key or the Smiles string [24].

After entering your search term (or browsing for all compounds), you will be presented with a list of relevant hits as shown in Figure 10 below.

![Figure 10 Compound list view.](image)

You can sort this list by ChEMBL [25] identifier [26] (which is the main way of identifying compounds in the diXa warehouse) or by compound name.

Clicking on the corresponding ChEMBL ID takes you to the compound details page (Figure 11).
Compound details page

The chemical compound details page (Figure 11) provides you with more information on the compound such as a full list of synonyms (searchable). It will also display a link to any disease datasets if the Comparative Toxicogenomics Database (see the disease section [27]) reveals an association between the chemical compound and the disease data for the liver, kidney and/or cardiovascular system.

Figure 11 Compound details page.

Additionally, you will find a link to the studies in the diXa warehouse which hold data on that specific chemical compound.

Additional chemical resources

The diXa warehouse provides you with links to other resources (ChEMBL and ChemAgora portal) if you are interested in exploring the chemical compounds in more detail. You can access the ChEMBL and ChemAgora page for a chemical compound from the compound list view (Figure 10) or the compound details page (Figure 11).
ChEMBL

ChEMBL [28] is a database of bioactive, drug-like small molecules [29] that contains 2D structures, calculated properties (e.g. molecular weight and Lipinski’s rule of five [30]) and abstracted bioactivity data (e.g. binding data and IC50 [31]) from multiple primary scientific journals. The database covers about 30 years of compound synthesis and testing and also contains annotated FDA-approved drugs. An example of a ChEMBL compound card is shown in Figure 12.

Figure 12 ChEMBL compound report card.
Learn more about ChEMBL

If you would like to know more about ChEMBL have a look at the ChEMBL quick tour [32].

The ChemAgora portal

The ChemAgora portal [33] is a chemicals database portal developed by the diXa project, which allow you to investigate chemical compounds (Figure 13). The portal holds heterogeneous data from selected third party systems and makes them available through a homogenous interface. You can access the ChemAgora portal through the diXa warehouse or directly through their website. The connection from the diXa warehouse to the portal is through the use of the standard InChI key identifiers which allows unequivocal identification of a chemical substance.

Figure 13 ChemAgora portal.
You can perform the following types of searches in ChemAgora:

- **Identifier** [26] based search (e.g. InChI key, CAS Register Number);
- Structure search: the portal carries out a search using a structure that you have drawn;
- Similarity search: the portal performs a search for compounds similar to a compound that you have identified through an identifier or a structure.

### How to submit data

#### Submitting via ArrayExpress

The easiest way to submit data to the diXa data warehouse is to submit your data via [ArrayExpress] [34]. The diXa project takes advantage of the existing structure used by ArrayExpress to archive functional genomics data.

#### Submitting directly to the diXa warehouse

Alternatively, you can submit your data directly to the diXa warehouse. Details explaining how to format your data for submission can be found [here] [35].

If you would like to submit your data directly to the warehouse then we recommend that you firstly get in touch with our *ddw* [at] dixa-fp7.eu (curator).

### Summary

- The diXa data warehouse is a curated collection of toxicogenomics experiments with cross-links to external resources such as chemical/toxicological and human disease databases.
- Making this data available to scientists will facilitate the building of computational models which in time could replace current animal-based test models for chemical safety.
- You can search or browse the diXa warehouse to explore toxicogenomics studies and see compounds used and analysis performed in these studies.
- The toxicogenomics studies in the warehouse are also complemented by publically available disease datasets.

### Quiz: diXa quiz

Try taking our short quiz - each question is a little case study of information you can find using the [diXa data warehouse] [36]. The quiz is scored automatically and you can repeat it as many times as you wish.

<table>
<thead>
<tr>
<th>Questions:</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempts allowed:</td>
<td><em>Unlimited</em></td>
</tr>
<tr>
<td>Available:</td>
<td><em>Always</em></td>
</tr>
<tr>
<td>Pass rate:</td>
<td>75 %</td>
</tr>
</tbody>
</table>
Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Find out more

Learn more about the diXa project on our website [13]. You can also access more diXa-related training materials from our training page [37].

Recommended courses

Below is a list of other relevant Train online courses:

- ArrayExpress quick tour [38] and tutorial [39]
- ChEBI quick tour [40]
- ChEMBL quick tour [32]

Getting in touch

If you have any questions or feedback you can get in touch with us from our contact page [41].

Links to external databases

The diXa warehouse search provides links to external chemical- and human-disease databases. These include:

- ChemAgora [33] - a chemical database portal (see the ChemAgora [42] section of this course);
- ChEMBL [28] - a database of bioactive, drug-like small molecules (see the ChEMBL [43] section of this course);
- ToxDb [44] - a toxicogenomics database that infers network effects of clinical toxins and drugs;
- Experimental Factor Ontology (EFO) [45] - provides a systematic description of many experimental variables available in EMBL-EBI databases and for external projects.
- ArrayExpress [12] - a database of functional genomics data;
- Comparative Toxicogenomics Database (CTD) [48] - a research tool which holds curated scientific data describing different relationships between e.g. chemical/drugs, genes/proteins, disease, etc.

References

1. Toxicogenomics [49] (Wikipedia page)


Contributors

Vera Matser [1]

EMBL-EBI
Training Liaison Officer - EBI Training Team

Source URL: https://www.ebi.ac.uk/training/online/course/dixa-data-warehouse-exploring-data-toxicogenomics

Links
[1] https://www.ebi.ac.uk/training/online/trainers/matser
[3] https://www.ebi.ac.uk/training/online/course/dixa-data-warehouse/references
[4] https://www.ebi.ac.uk/training/online/glossary/transcriptomics
[5] https://www.ebi.ac.uk/training/online/glossary/proteomics
[6] https://www.ebi.ac.uk/training/online/glossary/metabolomics
[8] https://www.ebi.ac.uk/training/online/glossary/toxicogenomics
[9] https://www.ebi.ac.uk/training/online/glossary/processed-data
[10] https://www.ebi.ac.uk/training/online/glossary/metadata
[11] https://www.ebi.ac.uk/training/online/glossary/function-genomics
[14] https://www.ebi.ac.uk/training/online/glossary/arrayexpress
[15] https://www.ebi.ac.uk/training/online/glossary/assay
[16] https://www.ebi.ac.uk/training/online/glossary/gene-expression
[17] https://www.ebi.ac.uk/training/online/glossary/principle-component-analysis
[18] https://www.ebi.ac.uk/training/online/glossary/profiles
[19] https://www.ebi.ac.uk/training/online/glossary/mirna
[20] https://www.ebi.ac.uk/training/online/glossary/methylation
[21] https://www.ebi.ac.uk/training/online/glossary/icd-10
[22] https://www.ebi.ac.uk/training/online/course/dixa-data-warehouse/compounds/compound-details-page
[23] https://www.ebi.ac.uk/training/online/glossary/synonym
[25] https://www.ebi.ac.uk/training/online/glossary/chembl
[26] https://www.ebi.ac.uk/training/online/glossary/identifier
[28] https://www.ebi.ac.uk/chembl/
[29] https://www.ebi.ac.uk/training/online/glossary/small-molecules
[30] https://www.ebi.ac.uk/training/online/glossary/lipinskis-rule-five
[31] https://www.ebi.ac.uk/training/online/glossary/ic50
[32] https://www.ebi.ac.uk/training/online/course/chembl-quick-tour
[34] http://www.ebi.ac.uk/arrayexpress/submit/overview.html
[38] https://www.ebi.ac.uk/training/online/course/arrayexpress-quick-tour
diXa data warehouse: Exploring data from toxicogenomics studies
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[39] https://www.ebi.ac.uk/training/online/course/arrayexpress-exploring-functional-genomics-data-ar
[40] https://www.ebi.ac.uk/training/online/course/chebi-quick-tour
[42] https://www.ebi.ac.uk/training/online/course/dixa-data-warehouse/compounds/exploring-chemical-compounds/chemagora-portal
[43] https://www.ebi.ac.uk/training/online/course/dixa-data-warehouse/compounds/exploring-chemical-compounds/chembl
[45] http://www.ebi.ac.uk/efo/
[46] http://www.ebi.ac.uk/gxa/home