Pacific Symposium on Biocomputing 12:169-180(2007)

SETUP X – A PUBLIC STUDY DESIGN DATABASE FOR METABOLOMIC PROJECTS

MARTIN SCHOLZ, OLIVER FIEHN[#]

University of California, Davis
Genome Center
451 E. Health Sci. Drive
Davis, California 95616, USA

Metabolomic databases are useless without accurate description of the biological study design and accompanying metadata reporting on the laboratory workflow from sample preparation to data processing. Here we report on the implementation of a database system that enables investigators to detail and set up a biological experiment, and that also steers laboratory workflows by direct access to the data acquisition instrument. SetupX utilizes orthogonal biological parameters such as genotype, organ, and treatment(s) for delineating the dimensions of a study which define the number of classes under investigation. Publicly available taxonomic and ontology repositories are utilized to ensure data integrity and logic consistency of class designs. Class descriptions are subsequently employed to schedule and randomize data acquisitions, and to deploy metabolite annotations carried out by the seamlessly integrated mass spectrometry database, BinBase. Annotated result data files are housed by SetupX for downloads and queries. Currently, 39 users have generated 48 studies, some of which are made public.

1. Metabolomic DBs require metadata on study designs

Metabolomic data can only be interpreted on the basis of background information of the experimental design of the biological parameters that were studied, and the details of data acquisition and data processing. Metabolites, unlike proteins, genes or RNA molecules, do not commonly carry specific information content. Instead, the role of metabolites in biological processes needs to be unravelled by their changes in levels, turnover rates and location in response to influence factors such as perturbation of the genetic constitution or external stress treatments.

Generally, cellular and organismal responses on such perturbations comprise many metabolic events. Only comparisons across a variety of biological studies and many different perturbation factors enable researchers to distinguish specific from unspecific effects, and therefore precisely define the meaning of metabolomic changes. Currently, no public metabolome database comprises

[#] Work supported by grant RO1 ES013932 of the NIEHS/National Institutes of Health.

such wealth of information on the actual conditions under which biological studies were carried out [1].

We are here presenting a solution for setting up metabolomic experiments, SetupX[a], comprising a description of the biological study design, a management of the experimental lifecycle and serving as a public database for metabolomic studies. The primary objective of this system is to capture the most relevant biological metadata for a study and to enable the user an easy access to upload or download and query such information. Secondary to its function as experimental metadata repository, SetupX directs the metabolite profiling data acquisition at the laboratory gas chromatograph-time of flight mass spectrometer and enables overview about scheduled experiments and data acquisition status. It serves as central interface for data processing tasks to the BinBase mass spectrometry database and for keeping result files for downloads.

SetupX therefore presents a fully functional and public database system integrating metabolomic workflows from conceptual design over laboratory practice to steering data processing tasks and result queries. We here detail the computational aspects of SetupX for reuse of metabolomic data sets for statistical analysis and cross-study investigations. Its functionality enforces researchers to carefully design and completely document biological studies.

2. Conceptualized Schema

SetupX has been developed over the past three years. Partly as result of work for the Food Standards Agency, UK[b], the first version of SetupX was based on the general ‘Architecture for a metabolomics experiment’ schema (ArMet) [2] that broadly classified the overall workflow and data facts into nine larger modules and relationships between these. In a similar manner to the later concept of MIAMET [3] (the minimal information on a metabolomic experiment), ArMet demands a description of the BioSource, the object and materialization of a biological study design. However, the internal structure and required ontologies supporting such BioSources descriptions remained vague and subject to the implementation of community-specific versions of ArMet. A similar vagueness of conceptual clarity and descriptive stringency was found in related omics areas, namely MAGE-ML [4] and proteomics database efforts. Most of the existing experimental design descriptors focused on the data

[a] SetupX. [http://setupx.fiehnlab.ucdavis.edu/m1/]

[b] Food Standards Agency: Safety Assessment of Genetically Modified Foods Research Programme (G02) [http://www.food.gov.uk/science/research/researchinfo/foodcomponentsresearch/novelfoodsresearch/g02programme/]

acquisition and processing parts rather than on the biological side of studies, which are indeed harder to describe and conceptualize. Promising efforts were presented by DOME, a database system for functional genomics and systems biology, covering various omic techniques[c]. The database schema embarks on thorough description of biological metadata defined by users; however, the underlying schema is still not universal enough to capture the breadth of study design in biology. The SetupX design therefore emphasizes the description of the BioSource and only demands pointers to documents for actual chemical processes used in sample handling and chemical preparations. Major efforts were reported by the SMRS group[d] which focused on biomedical and toxicological studies and whose work is now continued in the efforts of the Metabolomics Society.

Since SetupX was planned to house a very large set of different biological studies, spanning many disciplines from plant biology to clinical research, one of the most important tasks was to keep the schema adaptable to practical experiences and to ongoing discussions in each community with respect to organization and prerequisites of consistent and complete study design descriptions. At the same time, SetupX had to be flexible to account for a large variety of BioSources (spatial and genotypic descriptions of the physical objects that undergo metabolomic investigations, including their growth history) and treatments of these (experimental alterations of impact parameters influencing the metabolic states of BioSources). Consequently, SetupX utilizes a stringent

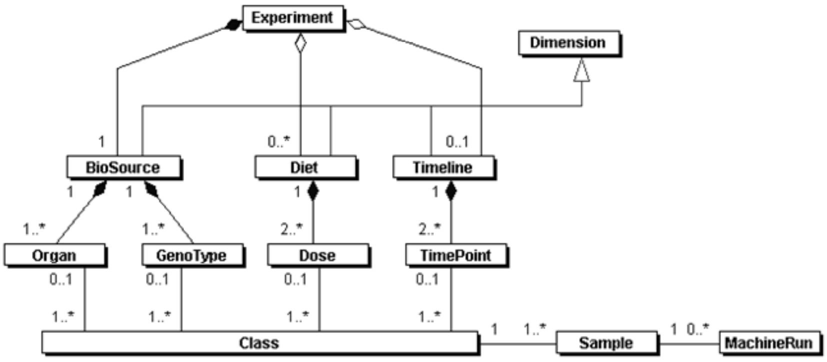


Figure 1 UML diagram of the experimental design

[c] DOME. [http://mendes.vbi.vt.edu/tiki-index.php?page=DOME]

[d] Standard Metabolic Reporting Structure [http://www.smrsgroup.org/documents/SMRS_policy_draft_v2.3.pdf]

schema that conceptualizes these two orthogonal properties of any BioSource, its physical object (including past environmental factors) and any process or parameter that investigators intentionally manipulated to enforce metabolic responses – the treatment (including time course and dose descriptors). Therefore, BioSources for metabolomic studies require complete descriptions of the object as well as any manipulation of the object that distinguishes it from related objects in the same study. Hence, BioSource and treatment define the most important vectors that span the dimensions of a metabolomic study, called classes. In principle, further dimensions can be spanned by using different chemical treatments or data acquisition methods, but mostly this is not intended in metabolomic studies. The number of these dimensions is not limited and varies according to the experimental design of each study. Objects that cannot be distinguished by any of the vectors are called bioreplicates and belong into the same class yet have unique object identifiers. Often, metabolomic data of these classes are later compared by statistical means in order to unravel metabolic effects that distinguish these classes. However, classes may also be combined to super-classes if certain distinguishing dimensions are deemed by investigators to be less important.

BioSources and treatments could thus comprise of any biological experiment and were not constricted to a certain experimental condition. The demands for such flexibility created a challenge for developing SetupX and to enable users an easy access to populate the database. SetupX has met this challenge by spanning the dimensions on the fly while users enter information that classifies distinct BioSource or treatment parameters. For example, each genotype, each organ, each cell type or each difference in age, sex or past growth location defines classes ('BioSource'), as well any intentionally altered parameter such as nutritional regimen, chemical elicitors or time lines that are imposed onto the

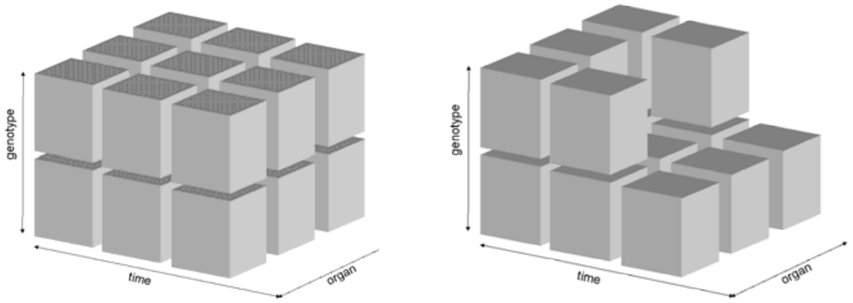
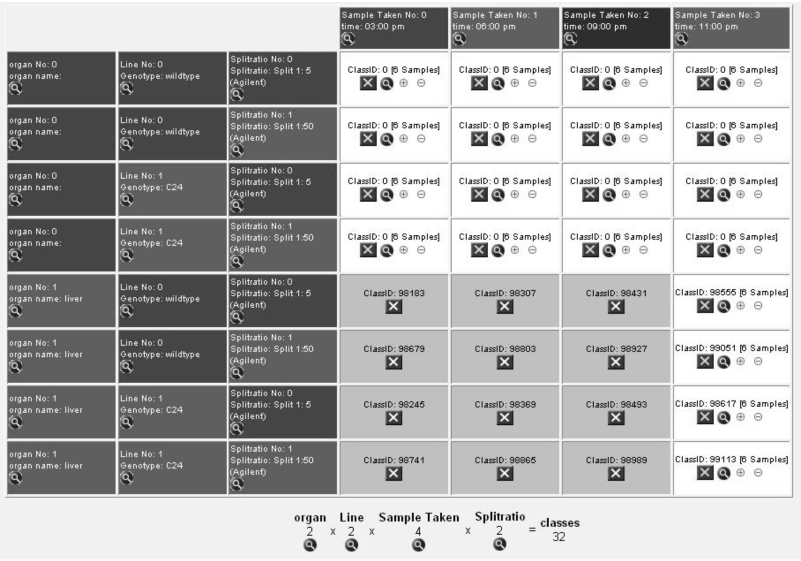


Figure 2. Three dimensional metabolomic study comprising 18 classes (2 genotypes x 3 organs x 3 time points, left panel) of which some classes may be void of bioreplicates and deselected in SetupX (right panel).

BioSources as part of the study (‘treatment’).

Figure 1 shows a simplified UML diagram of such an experimental design including the dimensions. Every class has a relation to one of the instances of variation of each dimension. Such study design can be conceptualized as cube if BioSource is distinguished by two vectors (genotype and organ) and treatment by one dimension (time) as shown in figure 2, left panel. The vector space represents the classes in the experiment – each possible combination of each variation per dimension represents one class, shown in the image a single cube. The maximal number *n* of classes spanned by *d* dimensions thus simply equals $n=\Pi d$. For specific studies, not every class may be populated by bioreplicates, i.e. not all dimensions may apply to all classes. For example, mice organs such as liver or kidney are usually studied after animals are sacrificed, whereas body fluids can be taken along treatment dimensions. In SetupX, users can therefore deselect certain classes that are void of bioreplicates.

Multidimensional designs cannot be easily displayed to users. Therefore, study designs in SetupX are visualized as table which can handle as many dimensions as needed (Fig. 3). Deselected classes are represented in grey shaded boxes.



			Sample Taken No: 0 time: 03:00 pm	Sample Taken No: 1 time: 06:00 pm	Sample Taken No: 2 time: 09:00 pm	Sample Taken No: 3 time: 11:00 pm
organ No: 0 organ name:	Line No: 0 Genotype: wildtype	Splitratio No: 0 Splitratio: Split 1: 5 (Agilent)	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]
organ No: 0 organ name:	Line No: 0 Genotype: wildtype	Splitratio No: 1 Splitratio: Split 1:50 (Agilent)	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]
organ No: 0 organ name:	Line No: 1 Genotype: C24	Splitratio No: 0 Splitratio: Split 1: 5 (Agilent)	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]
organ No: 0 organ name:	Line No: 1 Genotype: C24	Splitratio No: 1 Splitratio: Split 1:50 (Agilent)	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]	ClassID: 0 [6 Samples]
organ No: 1 organ name: liver	Line No: 0 Genotype: wildtype	Splitratio No: 0 Splitratio: Split 1: 5 (Agilent)	ClassID: 98183	ClassID: 98307	ClassID: 98431	ClassID: 98555 [6 Samples]
organ No: 1 organ name: liver	Line No: 0 Genotype: wildtype	Splitratio No: 1 Splitratio: Split 1:50 (Agilent)	ClassID: 98679	ClassID: 98803	ClassID: 98927	ClassID: 99051 [6 Samples]
organ No: 1 organ name: liver	Line No: 1 Genotype: C24	Splitratio No: 0 Splitratio: Split 1: 5 (Agilent)	ClassID: 98245	ClassID: 98369	ClassID: 98493	ClassID: 98617 [6 Samples]
organ No: 1 organ name: liver	Line No: 1 Genotype: C24	Splitratio No: 1 Splitratio: Split 1:50 (Agilent)	ClassID: 98741	ClassID: 98865	ClassID: 98989	ClassID: 99113 [6 Samples]

organ		Line		Sample Taken		Splitratio		classes
2	x	2	x	4	x	2	=	32

Figure 3 SetupX view of a four dimensional study.

3. Customization of dimensions by ontologies

There is no consensus in biology on the minimal, the necessary (required), the optimal or the maximal numbers of parameters that describe an acceptable study. Journals usually declare that materials and methods must be sufficiently described to understand and repeat a scientific report, but do not detail the parameters. Curation of reports is consequently performed in the peer-reviewing process that lacks consistent guidelines and thus leads to frustration among authors and reviewers. Database designs lack even this peer-reviewing process but must rely on automatic consistency checks. SetupX utilizes consistency checks on the level of dictionaries (spelling and minimal word/letter counts), controlled vocabularies and ontologies that define the parameter space for selecting dimensions. By using ontologies we can map the real name (for example of an organ) to a unique identifier taken from the ontology and thus enable queries that are comparing different objects by using unique identifiers instead of "strings" labelling the information. Hence, queries are independent from use of synonyms and may span across different levels of abstractions.

SetupX is equipped with a connector to OBO ontologies. Consistency is checked by relating a specific ontology repository to each input field in the front end. The check in the current version is a simple lookup if the term entered by a user is defined in the related ontology. The Ontologies used in SetupX are currently the plantbased structure ontology from the Plant Ontology Consortium, the Arabidopsis development ontology from the Arabidopsis Information Resource (TAIR) and the Human developmental anatomy ontology from the Medical Research Council Human Genetics Unit Edinburgh U.K.

In addition, SetupX has built-in validations that work as a spelling check. A service from Google checks the number of results that were found. If the number of results found is high enough, the value will be accepted, but it will be rejected if the returned number is lower than a defined minimum. Such simple validity checks can be assigned to any of the input fields in the system in order to prevent 'dummy' entries.

For example, selection of BioSource 'human' and a plant organ 'leaf' is disabled by SetupX using the powerful NCBI species taxonomy [5] for species definitions that informs the (subsequent) definition of organs that are selected for metabolomic studies. Use of the NCBI taxonomy enables queries for synonyms or generalized terms such as the genus 'rat' for any of the 23 rat species that are currently defined at NCBI.

Organ selections subsequently depend on the species under study. A good example is the definitions of organs and their relationships given in PlantOntology [6] that can directly be utilized within the SetupX schema.

However, such dependencies cause practical problems for clarity in use. For example, if two different species were selected in a single metabolomic study (e.g. ‘human’ and ‘soybean’), different subsequent views would result asking the user for input of specific parameters depending on the different biological ontologies that exist for these species. Whereas such different views can in principle be implemented, actual user access demonstrated that biologists might easily get confused by the number of required input parameters. Instead, SetupX enforces splitting such study into two independent experiments that can later be combined on the level of result downloads for data processing tasks and statistical comparisons.

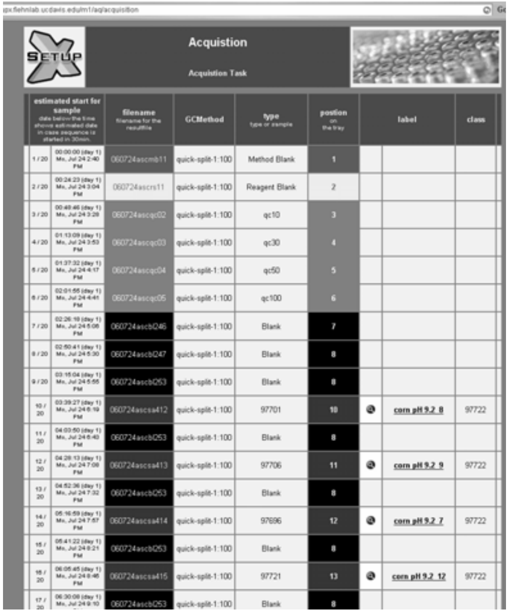
In addition to community accepted ontologies on the organ and genotype level, we have hard coded parameters further describing dimensions such as ‘past growth conditions’. The current version of SetupX supports parameter definitions for humans, animals, plants and mircoorganisms. This separation is obviously not scientifically exhaustive but instrumental for adjusting user interfaces for parameter input and keeping logic consistency of the database. Customized sets of (required or optional) parameter inputs were realized by using the taxonomic tree structure by navigating from species node up to the first match for a node that would classify all underlying nodes.

For example, if a selected species belongs to the plant kingdom, growth conditions on light, humidity, temperature, soil, location, developmental stage and others are requested, complying to the draft document of the minimal reporting standards requested to describe a plant metabolomic experiment which was recently released by the Metabolomics Society[e]. For the species ‘human’, a different set of parameters is requested such as gender, age, body mass index and others. Depending on the actual study, however, certain parameters need to be detailed to understand the study design and some of these parameters may even only be released long after a metabolomic experiment is finished such as ‘survival rate’ for cancer studies. Hence, maintenance of logic consistency of such a database is an ongoing challenge due to the huge number of parameters and study types that may influence the metabolic phenotypes. In a similar way like journals, SetupX asks study investigators to detail as many parameters as possible but does not comprise many required fields. Instead, documents detailing further parameters for a given study may be uploaded by investigators. Such documents will inform the further development of SetupX hard coded parameter fields.

[e] Metabolomics Standards Initiative [http://msi-workgroups.sourceforge.net/]

4. Study design classes steer laboratory workflows

A given metabolomic study may comprise many classes and even more bioreplicates. Based on experience and simple statistical considerations, the minimal number of bioreplicates populating a class is set as six in SetupX, whereas the optimal number of bioreplicates per class depends on a power analysis that takes the natural variability of metabolic levels into account. This variability is much higher in uncontrolled situations such as human (cohort) studies than under controlled laboratory conditions utilizing near-homozygous genotypes and specific nutritional regimes. Consequently, small metabolomic studies typically comprise some 48 bioreplicates whereas larger studies easily contain hundreds, sometimes thousands of bioreplicates. The largest study included in SetupX is a project on 12 potato genotypes x 4 field trial growth locations, each class populated with 28-30 bioreplicates which totals to more than 1,300 samples. This study was funded by the British Food Standards Agency in 2003 in order to test substantial equivalence of genetically modified and classically bred potato tubers, and result data sets for a field trial using the same experimental design (but under different environmental conditions for year 2001) has previously been published.



px.fiehnlab.ucdavis.edu/m1/aq/acquisition

Acquistion

Acquistion Task

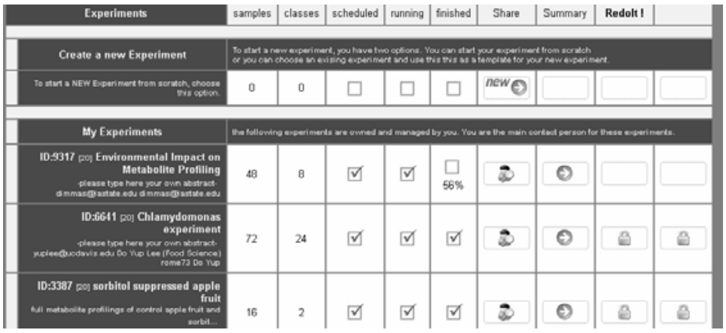
estimated start for sample date below the time shows estimated date in case sequence is started in 30min.		filename filename for the resultfile	GCMethod	type type or sample	position on the tray	label		class
1 / 20	00:00:00 (day 1) Mo, Jul 24 2:40 PM	060724ascmb11	quick-split-1:100	Method Blank	1			
2 / 20	00:24:23 (day 1) Mo, Jul 24 3:04 PM	060724ascrs11	quick-split-1:100	Reagent Blank	2			
3 / 20	00:48:46 (day 1) Mo, Jul 24 3:28 PM	060724ascqc02	quick-split-1:100	qc10	3			
4 / 20	01:13:09 (day 1) Mo, Jul 24 3:53 PM	060724ascqc03	quick-split-1:100	qc30	4			
5 / 20	01:37:32 (day 1) Mo, Jul 24 4:17 PM	060724ascqc04	quick-split-1:100	qc50	5			
6 / 20	02:01:55 (day 1) Mo, Jul 24 4:41 PM	060724ascqc05	quick-split-1:100	qc100	6			
7 / 20	02:26:18 (day 1) Mo, Jul 24 5:06 PM	060724ascbl246	quick-split-1:100	Blank	7			
8 / 20	02:50:41 (day 1) Mo, Jul 24 5:30 PM	060724ascbl247	quick-split-1:100	Blank	8			
9 / 20	03:15:04 (day 1) Mo, Jul 24 5:55 PM	060724ascbl253	quick-split-1:100	Blank	8			
10 / 20	03:39:27 (day 1) Mo, Jul 24 6:19 PM	060724ascsa412	quick-split-1:100	97701	10	●	corn pH 9.2_8	97722
11 / 20	04:03:50 (day 1) Mo, Jul 24 6:43 PM	060724ascbl253	quick-split-1:100	Blank	8			
12 / 20	04:28:13 (day 1) Mo, Jul 24 7:08 PM	060724ascsa413	quick-split-1:100	97706	11	●	corn pH 9.2_9	97722
13 / 20	04:52:36 (day 1) Mo, Jul 24 7:32 PM	060724ascbl253	quick-split-1:100	Blank	8			
14 / 20	05:16:59 (day 1) Mo, Jul 24 7:57 PM	060724ascsa414	quick-split-1:100	97696	12	●	corn pH 9.2_7	97722
15 / 20	05:41:22 (day 1) Mo, Jul 24 8:21 PM	060724ascbl253	quick-split-1:100	Blank	8			
16 / 20	06:05:45 (day 1) Mo, Jul 24 8:46 PM	060724ascsa415	quick-split-1:100	97721	13	●	corn pH 9.2_12	97722
17 / 20	06:30:08 (day 1) Mo, Jul 24 9:10	060724ascbl253	quick-split-1:100	Blank	8			

Figure 4: SetupX data acquisition task download

Typical cycle times for an individual sample per metabolomic data acquisition is about 30 min, or about 40 samples per day plus quality control samples. Data acquisition instruments show drifts in sensitivity and resolution, especially mass spectrometry based technology platforms. In order not to bias statistical analyses or the metabolomic data structure by non-biological factors such as machine drift, bioreplicates (classes) need to be randomized across the whole data acquisition sequence. In addition, each sample needs to unambiguously match the unique bioreplicate identifier in SetupX. Laboratory staff downloads the randomization schemata, sample pre-treatment methods (such as ‘extraction protocols’) and data acquisition method parameters (such as ‘split ratio’ or ‘detector voltage’) directly from SetupX, thereby limiting systematic or gross errors such as misspelling of file identifiers or misplacing samples. Importantly, use of different methods for sample pretreatments or data acquisition routines also generates new dimensions for class definitions. Consequently, sample preparation and instrumentation parameter differences are treated in the same manner like differences between biological parameters (BioSource or treatment dimensions). A square root blocking schema randomizes samples across data acquisition schedules, adding quality controls and blank control samples as mandatory part of the overall study (fig. 4).

Two partnering laboratories currently use SetupX, the UC Davis Genome Center’s metabolomics research and the metabolomics core laboratory, each with different laboratory staff and data acquisition machines. Raw data result files are processed and exported by post-acquisition macros. SetupX web interfaces enable investigators to keep track of the acquisition status by automatically checking for result file outputs and by reading the machine generated log files. Figure 5 shows personalized tracking information for three different experiments, two of which were completed while one of the experiments displayed a 56% completion status. SetupX is based on a modular design and can easily adapt to laboratory environments other than the current use of Leco’s gas chromatography/time of flight mass spectrometers.

Once complete, initial mass spectral result files are scheduled using a further SetupX GUI for users (or laboratory staff) to start subsequent data processing and metabolite annotation using the seamlessly integrated, but independent BinBase database[7]. BinBase receives information about the samples including the class structure which is essential for the calculation and starts an automatic



Experiments	samples	classes	scheduled	running	finished	Share	Summary	RedoIt !	
Create a new Experiment	To start a new experiment, you have two options. You can start your experiment from scratch or you can choose an existing experiment and use this this as a template for your new experiment.								
To start a NEW Experiment from scratch, choose this option.	0	0	□	□	□	new ➔			
My Experiments	the following experiments are owned and managed by you. You are the main contact person for these experiments.								
ID:9317 [20] Environmental Impact on Metabolite Profiling -please type here your own abstract- dimmas@iastate.edu dimmas@iastate.edu	48	8	☑	☑	□ 56%	👤	➔		
ID:6641 [20] Chlamydomonas experiment -please type here your own abstract- yuplee@ucdavis.edu Do Yup Lee (Food Science) rome73 Do Yup	72	24	☑	☑	☑	👤	➔	🔒	🔒
ID:3387 [20] sorbitol suppressed apple fruit full metabolite profilings of control apple fruit and sorbit...	16	2	☑	☑	☑	👤	➔	🔒	🔒

Figure 5 UML diagram of the experimental design

annotation. Filtered and annotated result data sets are reported back from BinBase and uploaded to Setup X from which investigator or the public can query both experimental metadata and metabolomic results.

SetupX protects access rights to specific experiments in different access levels, from reading (level 20) to modification (level 40), download (level 45) and experiment deletion rights (level 85). Only a few studies are currently publicly available, depending on publication of the major conclusions in peer-reviewed journals. Currently, SetupX details 48 studies comprising 4,500 samples with access rights for 39 users.

5. Implementation

SetupX has been developed as a server side application in the J2EE[f] framework using a relational database management system. It can therefore be installed on any certified J2EE application server. The flexibility of the system posed challenges for implementing the front end because the underlying schema may be subject to changes, with subsequent needs for front end adaptations. Therefore the user front end is generated using a combination of Java Server Pages (JSP) and Java Servlets. Attempts were unsuccessful to implement a user friendly functionality for capturing experimental designs by generating the front end based on an XML schema[g] as had been exemplified previously for PEDRo, a proteomics experiment database [8]. PeDRo inhibits an easy customization of front ends that are intuitive for first time users.

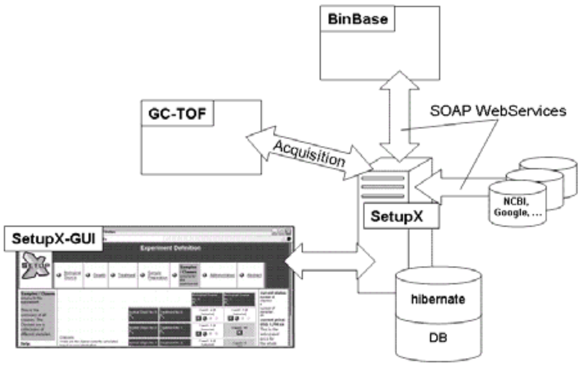


Fig 6. Communication via WebServices between SetupX and other services.

Most connections form SetupX to different services and databases are implemented by using SOAP WebServices[h].

WebServices enable communication between systems

[f] Java 2 Platform, Enterprise Edition [http://java.sun.com/javaee/]

[g] W3C XML Schema [http://www.w3.org/XML/Schema]

[h] W3C Web Services [http://www.w3.org/2002/ws/]

independently of their implementation and programming language by using XML as a self describing exchange format. Most importantly, SetupX is integrated with the metabolic annotation database BinBase by using WebService as interaction technology. Figure 6 demonstrates how the system communicates with other services.

Each user request invokes several queries on the NCBI database, However, response times of the NCBI Service are too slow to be used in a live system. Therefore, the connection which was implemented as a WebService (Entrez[i]) was replaced by mirroring the whole NCBI taxonomy database locally with weekly downloads of updates. Based on these experiences, other resources were handled in the same manner by installing local copies such as the description of plant organs in Plantontology.org.

6. Conclusions and Challenges

SetupX presents a conceptually clear and pragmatic implementation of a database solution to set up and describe metabolomic studies, from study design to laboratory workflow and result data housing. Most importantly, it encompasses biological metadata to enable the public to reuse metabolomic data sets and to gradually learn more about specific versus unspecific metabolic responses to study parameters like ‘abiotic stress’. SetupX empowers queries across studies such as ‘Which experiments are present for a certain species?’ or ‘Download all data corresponding to plant leaf studies.’ Data can also be queried from the perspective of metabolites such as ‘Report all data referring to a specific compound.’ Obviously, such queries yield more interesting results with a growing numbers studies stored in the system.

However, community efforts such as the Metabolomics Standards Initiative are needed to further define minimal (required) and optimal (best practice) reporting standards. The database schema employed here can easily be replaced by a different schema or mapped onto other schemas once consensus formats are established by biological communities. Such consensus schemas will truly enable exchange of studies just by the transfer of a file, not by parsing the contents of scientific reports or by repeating studies. The difficult part here is to convince biologists to undergo the efforts to carefully populate the study design databases. We envision an intelligent import of experimental metadata from a range of typical document types such as Excel sheets by automatically analyzing the document data structure and recreating a blueprint of that particular study in SetupX.

[i] Entrez [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html]

The variability of biological study designs is so great that it is hard to imagine devising a perfect way to painlessly parse all relevant biological metadata from documentations held by investigators. Two examples may highlight these challenges: (a) Pharmacological studies frequently involve rodents. However, the genetic repertoire of standard laboratory mouse strains is not reflected by NCBI species codes but differs between individual laboratories or suppliers, based on a complex progeny and breeding schema. (b) Clinical studies challenge the presented study design in yet another, very different way. Classes may be compiled from a variety of patient (or volunteer) data, due to the individuality of every human subject that reflects the corresponding unique genotypic, phenotypic and societal context. In addition, a number of diseases are too rare to collect a high enough number of specimen for thorough statistical treatments. Despite great efforts to match patient and control subjects, often metadata that are acquired in follow up studies justify to regroup subjects into different study design classes or to carry out other ways of statistical analyses. It is therefore very hard to accurately represent the wealth of clinical patient metadata that could potentially impact metabolic phenotypes and simultaneously to keep strict patient privacy.

References

[1] Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Steinhauser D, et al. *Bioinformatics* 21:1635-1638 (2005).

[2] Jenkins H, Hardy N, Beckmann M, Draper J, Smith A, Taylor J, Fiehn O, Goodacre R, Bino R, et al. *Nature Biotechnology* 22:1601-1606 (2004).

[3] Bino R.J., Hall R.D., Fiehn O. et al. *Trends Plant Sci.* 9 418-425 (2004).

[4] Spellman P, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, et al. *Genome Biol* 3:research0046.1-0046.9. (2002).

[5] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, et al. *Nucleic Acids Res.* 1;34:D173-80 (2006).

[6] Jaiswal P, Avaraham S, Ilic K, Kellogg E, McCouch S, Pujar A, Reiser L, et al. *Comp Func Gen* 6:388-397 (2006)

[7] Fiehn O, Wohlgemuth G, Scholz M. *Proc. Lect. Notes Bioinformatics* 3615, 224-239 (2005)

[8] Garwood K, McLaughlin T, Garwood C et al. *BMC Genomics*. 5: 68 (2004)