doi: 10.1111/j.1365-3040.2009.01964.x

Xeml Lab: a tool that supports the design of experiments at a graphical interface and generates computer-readable metadata files, which capture information about genotypes, growth conditions, environmental perturbations and sampling strategy

JAN HANNEMANN[1,2], HENDRIK POORTER[3], BJÖRN USADEL[1], OLIVER E. BLÄSING[1,4], ALEX FINCK[1], FRANCOIS TARDIEU[5], OWEN K. ATKIN[6], THIJS PONS[2], MARK STITT[1] & YVES GIBON[1,7]

[1]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Golm, Germany, [2]University of Victoria, Centre for Forest Biology, PO Box 3020 STN CSC Victoria, Canada BC V8W 3N5, [3]Institute of Environmental Biology, Utrecht University, PO Box 800.84, 3508 TB Utrecht, The Netherlands, [4]Metanomics GmbH, Tegeler Weg 33, D-10589 Berlin, Germany, [5]INRA, Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux, 2 Place Viala, F-34820 Montpellier, France, [6]Ecosystem Dynamics Group, Research School of Biological Sciences, Australian National University, Canberra, ACT 2602, Australia and [7]INRA-Bordeaux, Université de Bordeaux, UMR619 Biologie du Fruit, 71 Avenue Edouard Bourlaux, F-33883 Villenave d'Ornon, France

ABSTRACT

Data mining depends on the ability to access machine-readable metadata that describe genotypes, environmental conditions, and sampling times and strategy. This article presents Xeml Lab. The Xeml Interactive Designer provides an interactive graphical interface at which complex experiments can be designed, and concomitantly generates machine-readable metadata files. It uses a new eXtensible Mark-up Language (XML)-derived dialect termed XEML. Xeml Lab includes a new ontology for environmental conditions, called Xeml Environment Ontology. However, to provide versatility, it is designed to be generic and also accepts other commonly used ontology formats, including OBO and OWL. A review summarizing important environmental conditions that need to be controlled, monitored and captured as metadata is posted in a Wiki (http://www.codeplex.com/XeO) to promote community discussion. The usefulness of Xeml Lab is illustrated by two meta-analyses of a large set of experiments that were performed with *Arabidopsis thaliana* during 5 years. The first reveals sources of noise that affect measurements of metabolite levels and enzyme activities. The second shows that *Arabidopsis* maintains remarkably stable levels of sugars and amino acids across a wide range of photoperiod treatments, and that adjustment of starch turnover and the leaf protein content contribute to this metabolic homeostasis.

Key-words: bioinformatics; data management; data mining; ontology.

Correspondence: J. Hannemann. e-mail: jaha@uvic.ca. Y. Gibon. e-mail: ygibon@bordeaux.inra.fr

INTRODUCTION

The sequencing of plant genomes together with the development of 'global' techniques to evaluate phenotypic traits at different levels, such as transcriptomics, proteomics, metabolomics and whole plant imaging, are leading to a shift from reductionist to more holistic approaches in plant research. This is facilitated by repositories in which microarray and other profiling data are collated from large numbers of published experiments, quality checked and made publicly available (Stoeckert, Causton & Ball 2002; http://affymetrix.arabidopsis.info/; Genevestigator, https://www.genevestigator.ethz.ch/). These repositories enormously increase the amount of data that can be accessed to extract new information without needing to perform additional experiments, and to support and extend the interpretation of new data sets. Data mining to extract relevant information from complex data sets (Frawley, Piatetsky-Shapiro & Matheus 1991) is becoming one of the most important tools in biology (Thimm *et al.* 2006). Unravelling the complexity behind phenotypes relies heavily on achievements in bioinformatics, including data management, integration and interpretation. It also requires standardized conceptualizations with explicit specifications (Gruber 1993). Standardization is not only needed for the phenotypic data; it is also important to have standardized descriptions of the genetic material, the growth conditions and the experimental design. The latter is increasingly referred to as 'metadata' (data about data).

The generation of high-quality molecular and physiological data depends strongly on the experiment being correctly designed, and interpretation of the data depends on the experiment being adequately described. Plant phenotypes typically result from complex interactions between the genotype and a highly dynamic environment. Good examples are the acquisition of freezing tolerance or thermotolerance following the exposure to a mild low temperature or heat stress, respectively (Atkin *et al.* 2006; Kotak *et al.* 2007). An even more striking example is the vernalization response, in which the regulation of genes induced by low temperatures is inherited through successive mitotic divisions and is even transmitted to the next generation (Dennis & Peacock 2007). Furthermore, it is not trivial to change the environment in a specific manner, because changing one feature of the environment may lead to secondary effects. For example, under high light, plants grow faster and, transpire more, which makes them more vulnerable to nutrient deficiency or water shortage. Unfortunately, the experimental design, the state of the plant material and the growth conditions, are often poorly defined. This is partly because experiments are not always precisely described. Even when the information is available, it is extremely time-consuming to manually extract it from written papers or databases.

Several initiatives have been undertaken to impose controlled and machine-readable vocabularies and/or ontologies for the annotation of genes (Berardini *et al.* 2004) and germ plasm (McLaren *et al.* 2005), the documentation of plant anatomy and developmental stages (Jaiswal *et al.* 2005; Ilic *et al.* 2007) and the description of experiments (Brazma *et al.* 2001; Bino *et al.* 2004; Zimmermann *et al.* 2005; see also the EnvO (for Environment Ontology) project at http://darwin.nerc-oxford.ac.uk/gc_wiki/index.php/EnvO_Project). By definition, controlled vocabularies simply consist of lists of terms, while ontologies provide a structured terminology that is implemented with precise specifications about the terms and their use, and which requires a minimal commitment to support sharing activities (Gruber 1993).

The description of growth conditions is an especially large challenge. There is considerable theoretical and empirical knowledge about how to grow plants. Requirements for light, temperature, water, nutrients and other environmental conditions have been studied in detail in a range of model species (e.g. in Ingestad 1982). Various efforts have also been made to standardize the description of environmental conditions (Krizek 1982; Langhans & Tibbitts 1997). However, to our knowledge, this information is rather scattered, and the validity of the existing vocabularies and ontologies is a matter of debate (Hastings, Lalloo & Khoo 2006). Firstly, they often consist of subjective guidelines, and the level of detail that must be reported is open to interpretation (Edgar & Barrett 2006). Secondly, to our knowledge, there is no existing machine-readable system addressing specific needs in plant physiology, i.e. the possibility to describe precisely physical or chemical variables such as light intensity and concentrations of nutrients. Available systems propose terms consisting of situations, for example, treatments or habitats, which often result from numerous rather than single variables. In this context, a major challenge is the need to describe the temporal dynamics. Changes in the environment can occur once or repetitively, regularly or irregularly, and gradually or abruptly, with different consequences for the phenotype. Experiments often involve changes of one or more features of the environment, sometimes at multiple time points. This information needs to be captured in the description of the growth conditions. To allow machine reading, it will be necessary to record traits and environmental conditions as a clearly defined function of time, rather than as a fixed descriptor and/or as free text. The use of thermal time to model the effect of temperature on plant development or growth rate is a classic example indicating the usefulness of such a strategy (Granier & Tardieu 1998; Sadok *et al.* 2007). This example also illustrates added value that experimenters can extract by including metadata about the growth and experimental conditions in their data analysis.

In addition to growth conditions, the description of biological experiments requires information about the genotypes, the developmental stage of the plants, and the nature of the organs being observed. Some local information (experimenter's names, IDs of growth chambers) may also be collected. Ideally, such a multilevel description eventually involving several ontologies would be possible within one tool. Such a platform should also allow the choice of the most appropriate ontology, or at least subsets of a given ontology, depending on the context.

Such multiple constraints might be overcome by collecting, organizing and storing metadata in formats like eXtensible Mark-up Language (XML)-formatted text, which benefit from standard parsers that are available on all major platforms. As XML does not require any specific IT infrastructure or software, it can be used without fastidious setups (e.g. installation of central servers and databases). XML files can nevertheless be readily processed by modern database management systems such as Oracle or MS-SQL, if further data management is needed. So-called dialects can be derived from XML by defining specific schemes, which offer an environment to describe the structure and constrain the content. After defining the dialect, the programming libraries can be written to generate dialect-compliant files. Frameworks can be implemented to develop tools that exploit the dialect, for example, ontology editors and providers, data entry interfaces, data analysers or data providers linking the data in XML format with other types of data. This information technology tool could provide a promising solution to handle the complex metadata describing growth conditions, and to link them with germ plasm information as well as, for example, microarray data for which dedicated data storage strategies are already available (Barrett *et al.* 2007).

The largest challenge to capturing metadata is, however, probably the effort required from individual scientists, who have to obtain and enter the information into a suitable format. For example, while XML-derived languages provide many advantages, their use in the native format is fastidious and impractical without bioinformatics support. It is therefore important to provide user-friendly interfaces that generate and manage the files while storing the data in the background (see e.g. Rayner *et al.* 2006).

In the present article, we describe Xeml Lab, a generic platform that helps the user to plan experiments, and concomitantly generates metadata files. It provides a standardized graphical format to describe genotypes, growth conditions, experimental design and sampling, and to capture and link the metadata in a machine-readable form. We also present two illustrative applications of data mining using metadata provided via Xeml Lab. In a related Wiki, we provide a short review highlighting how environmental conditions can be measured and controlled, and pointing out potential pitfalls when an environmental factor is used as a treatment to study plant performance. This is planned to act as a forum where the community can discuss how environmental conditions can be controlled and measured, and the information captured as metadata.

MATERIAL AND METHODS

XEML language

An XML-schema definition file (xsd) was created to specify the vocabulary and structure of XEML. This schema is open and available at http://www.codeplex.com/XEML/. The development was realized using Visual Studio 2005 (Microsoft).

Framework

Xeml Framework was programmed in C# with Visual Studio 2005 and implemented with .Net 2.0 framework running on Windows XP and Vista environments (Microsoft).

The framework is composed of three main libraries: (1) Xeml Core, which provides the basic functionality to load Xeml documents, ontologies, sample providers and interfaces to integrate custom components; (2) Xeml Visualisation, which provides methods to visualize the storyboard of a given Xeml document; and (3) Xeml Utilities, which provides some user-friendly methods to apply complex tasks within XEML.

The framework is available for all major platforms. On Microsoft platforms, .Net Framework 2.0 is required. On all other platforms, Mono framework, which is an open source implementation of the .Net framework, can be used.

Xeml Framework is available as an open source project for programming purpose and can be downloaded freely at http://www.codeplex.com/XEML.

Environment ontology

Xeml Environment Ontology was created with Visual Studio 2005 in XML format by defining an XML-schema. The ontology was converted into the standard OWL and OBO formats to facilitate interchange with the scientific community. At present, updates made to the ontology in one of these formats can be integrated back (in a semi-automated fashion) into the native data format supported by Xeml Lab. The schema is available as an xsd file at http://www.codeplex.com/XeO/Release/ProjectReleases.aspx?ReleaseId=19143, the recent version of the ontology at http://xeml.mpimp-golm.mpg.de/ontologies/recent.

Interactive designer

Xeml Interactive Designer was programmed in C# and implemented with .Net 2.0 framework. Besides visualization features of the framework, KryptonToolkit (ComponentFactory) was used to develop the user interface. Xeml Interactive Designer is available for Windows 2000, XP and Vista and can be downloaded freely at http://xeml.mpimp-golm.mpg.de/ClickOnce/XiD_RC1/. The installation occurs via the ClickOnce technology (Microsoft) and requires Internet Explorer (Microsoft) or FireFox implemented with the FireFox ClickOnce add-on. The .Net Framework 2.0 is also required. In case it is missing, a bootstrap installer will be provided automatically, but in this case, administrative privileges will be required. The Designer is also available as an open source project at http://www.codeplex.com/XiD.

Getting started

Once finished, the installation program launches Xeml Interactive Designer automatically with a default experiment design. Updates will be made available automatically and a prompt of acceptance or rejection will be loaded when starting the program. At the first start of the Designer, a folder ‘templates’ is created within a folder ‘XemlStore’ which is located directly under ‘My Documents’. The former may be used to save Xeml files that can be reused as templates, for example, when experiments are repeated. These paths can also be customized within the Designer, via the Options menu (Tools then Options). Six tabs are accessible: (1) ‘General’; (2) ‘Resources’; (3) ‘Design’; (4) ‘Validation’; (5) ‘Table’; and (6) ‘Report’. In addition, the native source code of the experiment can be viewed and edited via Tools/CodeEditor or by pressing F6. The first tab provides an interface in which general information about the experiment (author, aim of the experiment, keywords) is entered. The Resources tab is used to link data resources. By default, Xeml Environment Ontology and Plant Structure Ontology for flowering plants version 1.5 (Plant Ontology Consortium at http://www.plantontology.org) are automatically linked and loaded into the software. Alternatively, the user can specify the location of any ontology compatible with Xeml Lab.

Plant material, biochemical analysis and data processing

To conduct an illustrative data mining approach to detect unsuspected sources of noise, data were selected from a large collection of experiments conducted from 2002 to 2007, according to the following criteria: *Arabidopsis thaliana* Col-0 plants grown in Percival growth cabinets (CU-36 product line) and harvested as described in Bläsing *et al.* (2005), wild-type, 12 h/12 h day/night regime, 20 °C, total light intensity between 120 and 150 μmol m^{-2} s^{-1}, and relative humidity between 80 and 90%, harvest at the end of the day. Enzymes and metabolites were determined as in Gibon *et al.* (2004a,b). After inspection of the measured metabolites and enzyme activities and removal of aberrant (e.g. negative or out of range) values, missing values were replaced by the median of the respective metabolite or enzyme data. The resulting data set comprised of a total of 28 separate experiments, with measurements of 15 enzyme activities in 11 of them, and measurements of a set of nine metabolites in 23 of them (in six experiments, both metabolites and enzymes were measured). For convenience, the experiments are numbered in chronological order from 1 to 28.

For statistical analysis and visualization, data were centred to 0 mean, and scaled, i.e. divided by the root mean square. The resulting centred and scaled data matrices of either metabolites or enzymes were subjected to a principal component analysis (PCA; Pearson 1901), as implemented in R (R Development Core Team 2006). As the data were centred and scaled, PCA provides an estimator of the influence of different variables independently of the range of each variable. Colouring of data points was conducted using the year, the harvester or the identifier of the growth cabinet in which plants were grown.

To conduct an illustrative data mining approach to show how novel biological information can be extracted, further treatments and experiments were included in the abovementioned data set. Data obtained from wild-type Col-0 plants transferred to an extended night (Gibon *et al.* 2006; Usadel *et al.* 2008), and from wild-type and starchless *pgm* mutant Col-0 plants were obtained from previously published experiments (Gibon *et al.* 2004a,b, 2006; Thimm *et al.* 2004; Bläsing *et al.* 2005; Usadel *et al.* 2008). Data corresponding to plants grown at various photoperiods were obtained from unpublished experiments (Gibon *et al.* 2009). In the latter case, wild-type plants and *pgm* mutants were grown as described above, except that they were transferred to 5 h/19 h, 6 h/18 h, 7 h/17 h, 8 h/16 h, 9 h/15 h, 16 h/8 h and 20 h/4 h day/night regimes for 2–5 weeks. In addition, wild-type plants were also grown at 2 h/22 h, 3 h/21 h and 4 h/20 h day/night regimes and in exactly the same growth conditions. Sucrose and glucose were determined in ethanol extracts as in Jelitto *et al.* (1992), starch as in Hendriks *et al.* (2003) and total protein content as in Bradford (1976). Assays were prepared in 96-well microplates using a pipetting robot (Multiprobe HT, EP3 or Janus, Perkin-Elmer, Zaventem, Belgium). Absorbances were read at 340 or 595 nm in a Synergy, an ELX-800 or an ELX-808 microplate reader (Bio-Tek Friedrichshall, Germany). After data were checked for consistency and outliers such as negative values and non-replicated values which were inconsistent by an order of magnitude or more were removed, the resulting values were computed in R and displayed as boxplots to check for further inconsistencies (by story or time).

RESULTS AND DISCUSSION

Overall structure of Xeml Lab

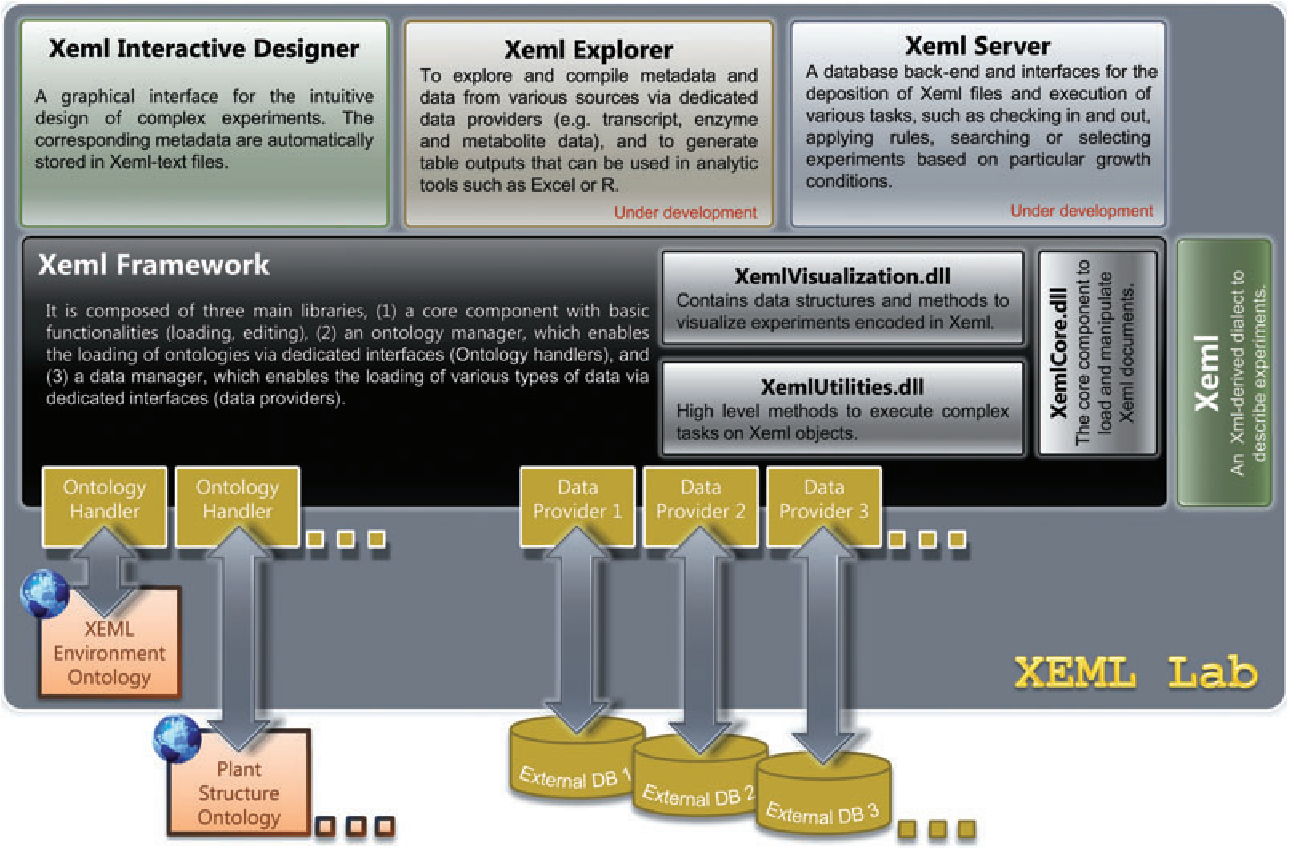
The documentation of biological experiments requires the capture of large amounts of heterogeneous data. This includes information about the genotypes used, quantitative or qualitative variables that describe the environment and any changes made in it during the experiment, and a description of the sampling strategy in terms of number of replicates as well as pooling and/or dissection of individuals. Xeml Lab has been created to assist plant biologists to design and document experiments in a machine-readable format, to link this metadata with the data generated in the corresponding experiments and, ultimately, to make both metadata and data available for data mining. It is flexible, extensible, as well as human- and machine readable.

The first version of Xeml Lab (Fig. 1) consists of a language, a framework to work with Xeml documents, develop tools and integrate custom modules, a user interface (Xeml Interactive Designer), a first working draft of an ontology of terms describing the abiotic environment (Xeml Environment Ontology), and providers to allow ontologies in other formats than XEML to be used. For each experiment, metadata describing (1) genotype information; (2) abiotic growth conditions as functions of time; and (3) sampling strategy are captured via an intuitive visual mode, and stored in an Xeml file that can be exported in table format. Extensions are planned that will allow metadata in the Xeml file to be automatically mapped to analytic data, which are stored in external files or databases (Fig. 1).

Flexible choice of ontologies

XEML can handle various libraries of terms and recommendations in parallel. This allows users to select the ontology of their choice, and to even use several alternative ontologies. Dedicated handlers and providers support the loading of ontologies, and give access to any type of data (e.g. stored in text files or in databases). As many ontologies are developed in OBO or OWL formats, we are developing providers that allow OBO and OWL files to be used in XEML. Thus, a provider enabling the use of Plant Ontology (see below) has already been written. Further details are given in the handbook at http://xeml.mpimp-golm.mpg.de/dnn/Resources/tabid/56/Default.aspx.

In the current version of Xeml Lab, the harvested plant material (organ, developmental stage, etc.) is described using Plant Ontology, which is available in OBO format (Jaiswal *et al.* 2005). Although several nomenclatures are currently being developed to describe growth conditions (see Introduction), we have developed a new environment ontology that is limited to the 'essential' environmental conditions but, importantly, is able to cope with temporal dynamics (see below). Genotype description would greatly benefit from an ontology dedicated to germ plasm information. Even though plants have been classified into taxa for centuries, to our knowledge, germ plasm ontologies are not yet available. This task is complicated by the high genetic



A graphical interface for the intuitive design of complex experiments. The corresponding metadata are automatically stored in Xeml-text files.

To explore and compile metadata and data from various sources via dedicated data providers (e.g. transcript, enzyme and metabolite data), and to generate table outputs that can be used in analytic tools such as Excel or R.

A database back-end and interfaces for the deposition of Xeml files and execution of various tasks, such as checking in and out, applying rules, searching or selecting experiments based on particular growth conditions.

It is composed of three main libraries, (1) a core component with basic functionalities (loading, editing), (2) an ontology manager, which enables the loading of ontologies via dedicated interfaces (Ontology handlers), and (3) a data manager, which enables the loading of various types of data via dedicated interfaces (data providers).

Contains data structures and methods to visualize experiments encoded in Xeml.

High level methods to execute complex tasks on Xeml objects.

The core component to load and manipulate Xeml documents.

An Xml-derived dialect to describe experiments.

Figure 1. Structure of the Xeml Framework.

diversity at the level of the species and subspecies, especially when genotypes result from intra- or interspecific crosses (hybrids) or backcrosses (introgression lines, recombinant inbred lines, near isogenic lines). We decided to leave this problem open at this time and, as a provisional step, to implement a 'stop-gap' solution in which species, accession, mutation and transformation event can be defined and commented on. Finally, XEML can be used in combination with specialized vocabularies or ontologies, for example, to facilitate the documentation of local information such as investigators names or IDs of growth chambers.

Xeml Interactive Designer

Reading and editing Xeml files with a text editor is difficult and time-consuming for non-specialized users. The key element of Xeml Lab is the Xeml Interactive Designer, which is a user interface dedicated to the design of biological experiments. It provides an interactive timeline visualization, which enables the intuitive creation and editing of the metadata that describe experiments while automatically generating Xeml files containing this information in the background. The following section describes the main features of the Xeml Interactive Designer. A more comprehensive help section and a tutorial are included in the software.

Designing and documenting an experiment

The Xeml Interactive Designer view tab (Fig. 2) is subdivided into three areas: a storyboard on the top, a story content on the left hand side and a variable definition area at the bottom. The story content currently includes two tabs, an environment ontology browser and a genotypes manager.

The storyboard provides the interface where the user develops and enters the experimental design. A story is a recipient for variables that describe along a continuous timeline the environment in which one or more individuals of one or more genotypes are growing. The storyboard enables the drawing of stories and derived stories (splits, basically this means a change in the conditions), and the positioning of events (which may be planned events, or documented unplanned accidents) and observation points (e.g. sampling times).

To start a story, two types of variables are defined using the story content section (Fig. 2). The first set is the genotypes. After activating the genotypes browser, genotypes are entered using the 'add a new genotype' button, using

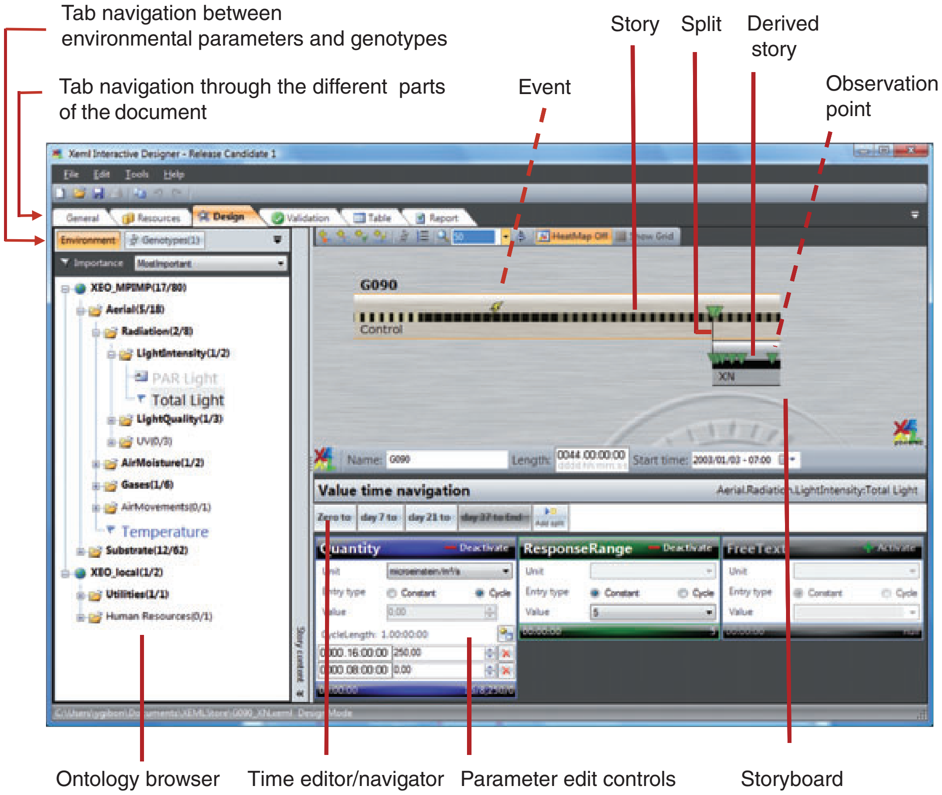
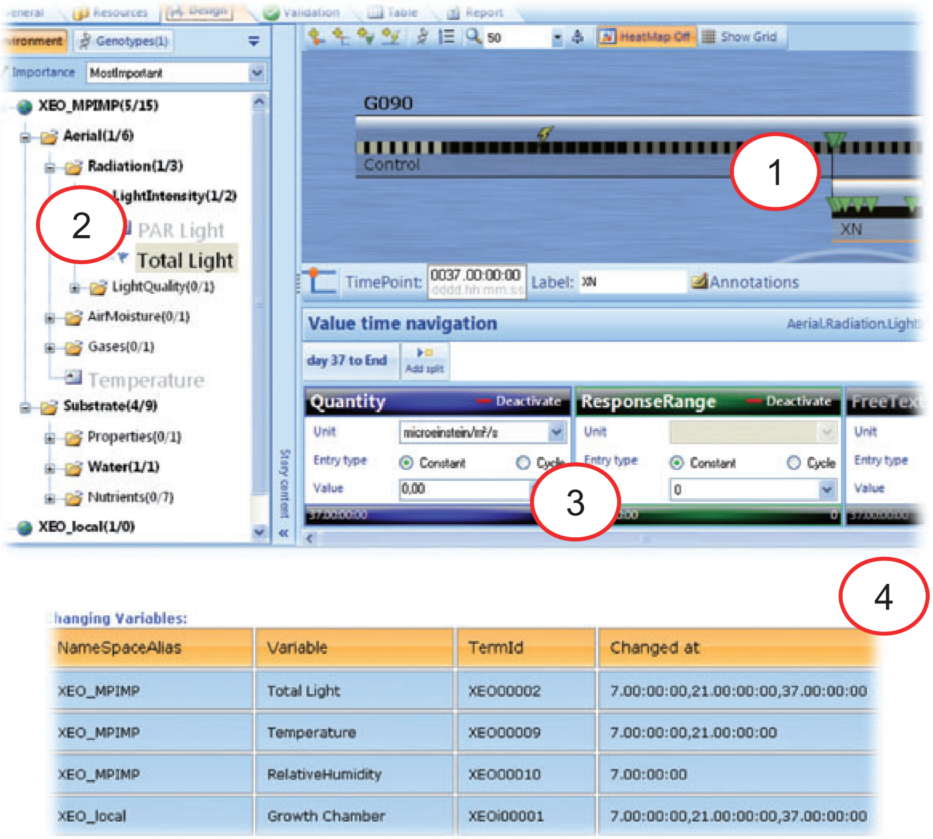


Figure 2. Xeml Interactive Designer – the Design tab. In this example, seeds were sown at 0700 h on the 3rd of January 2003 and the seedlings grown for 1 week in a 16 h/8 h day/night regime, total light intensity of 250 μmol m^{-2} s^{-1}. They were then transferred to an 8 h/16 h day/night regime, total light intensity of 100 μmol m^{-2} s^{-1}, for 2 weeks and pricked at day 14. At day 21, they were transferred to a 14 h/10 h day/night regime, total light intensity of 150 μmol m^{-2} s^{-1}. The 'heatmap' mode has been set to visualize the differences in light intensity. Then, at day 37 (split), some of the plants were transferred to continuous darkness (derived story 'XN'). Plants were harvested as indicated by the observation points, i.e. 10 min before the end of the night, then at 2, 6 and 12 h day ('Control'), and at 2, 6, 12, 24, 48 and 72 h following the transfer to continuous darkness ('XN'). The corresponding file (G090.xeml) is available from the supporting information.

editing tools (scroll, add, edit, delete) within the genotype edit form (not shown). The second set of variables defines the environmental conditions at the start of the experiment. After activating the environment ontology browser, a tree view shows the ontology terms organized according to the category or subcategory to which they belong (Fig. 2, left hand side). A filtering option enables selection of terms according to their importance. A click on a term loads the corresponding definition panel into the variable definition area; a double click initializes it with default values, which can then be modified as needed. As discussed later, values can be entered in three different context boxes (in the Xeml Environment Ontology: Quantity or Quality, ResponseRange and FreeText). It is important to note that variables are always defined as a function of time. This means that one or more time periods have to be set, within which a given variable is defined. For each time period, the parameter can be defined as constant, or as cycling. Time periods defined for the first variable will be inherited by default for the next variables that are entered but can be manually modified. Often, many of the conditions in a series of experiments are similar. To aid the user, a template folder, in which Xeml files describing routine growth conditions can be stored, is created automatically during the installation of the software.

At any time point within a story, it is possible to define a split and create a derived story, which inherits all information defined upstream, except for the characters that are explicitly modified in the derived story. This allows experimental manipulations that affect one or more of the environmental variables to be defined and entered via the graphical interface. In the storyboard of Fig. 2, plants were grown in an alternating 14 h light/10 h dark cycle (the story) and some were then transferred at the end of the night to continuous darkness (a derived story). Periods of light and dark are explicitly shown as light and dark regions along the storyline. The steps in generating a derived story are shown in more detail in Fig. 3.

Events (depicted as small yellow lightening-arrows) can be used to document unforeseen situations, or environmental conditions that are not covered by the ontology used (e.g. in the Xeml Environment Ontology, fungal infection, pesticide treatment or pricking).



Importance: MostImportant

50

TimePoint: 0037.00:00:00 dddd.hh:mm:ss Label: XN Annotations

Quantity: Unit microeinstein/m²/s; Entry type ● Constant ○ Cycle; Value 0,00

ResponseRange: Unit; Entry type ● Constant ○ Cycle; Value 0

hanging Variables:

NameSpaceAlias	Variable	TermId	Changed at
XEO_MPIMP	Total Light	XEO00002	7.00:00:00,21.00:00:00,37.00:00:00
XEO_MPIMP	Temperature	XEO00009	7.00:00:00,21.00:00:00
XEO_MPIMP	RelativeHumidity	XEO00010	7.00:00:00
XEO_local	Growth Chamber	XEOi00001	7.00:00:00,21.00:00:00,37.00:00:00

Figure 3. Story split. The derived story is defined by inserting a split at the appropriate time point, it inherits all information defined upstream (1); one or more variables that are changed in the derived story are selected from the ontology browser (2), and then modified in the variables definition area (3); every variable that has been modified will appear in the report tab (4). In this example, plants were grown under a 14 h/10 h day/night regime and were transferred to continuous darkness.

Observation points (depicted as small green triangles in Fig. 2) show the time points at which sampling occur. Thus, in the example shown in Fig. 2, samples are taken every 2 h during a photoperiod in control plants that were left in a 14 h light/10 h dark diurnal cycle, and at several times after transfer to continuous darkness. To define the sampling strategy, one or more observation points are selected by mouse click along the story and/or derived stories. The sampling wizard is then called up by mouse click. If more than one genotype is being grown, a list appears from which the user selects the genotypes that are to be harvested. The next step is to define the number of plants that will be pooled. The plant parts that are to be harvested (tissue, organ, whole individuals) are defined by choosing the appropriate term from the list provided by the ontology (for instance Plant Ontology). It can be specified if one individual is to be dissected into different parts, for example, roots and shoots. In the last step, the number of replicates is entered. It is possible to define a common sampling strategy for all of the genotypes, or individual sampling strategies (e.g. if one of the genotypes is much smaller and requires more individuals per sample). After confirmation, sample identifiers will be generated according to the story ID, and the time of harvest. It is possible to provide an estimate of the time that will probably be required to harvest one individual. This might be useful when large experiments are planned, to check if the sampling can actually be completed in an appropriate time span. For each observation point, a graphical overview of the sampling strategy can be viewed by clicking the link ‘n sample(s)’ in the variable definition area. Samples can also be erased individually from this area (‘clear samples’), or collectively from the Tools menu (‘clear all samples’).

Validation

A validation tab notifies the user about possible errors and/or omissions by displaying validation messages, which are collected from the framework. This step is critical, as ‘invalid’ Xeml files might be useless, especially for data mining purpose. However, the user is not forced to fix errors immediately, as this would make the use of the program uncomfortable. It is possible to save experimental designs with errors or omissions, so that they can be fixed later.

There are three levels of validation: (1) the format and the structure of the Xeml document itself; (2) the usage of ontology terms, for example, it is checked that all of the strongly recommended environmental variables are specified; and (3) specific modules that can be implemented and loaded by the framework to extend the validation logic and to apply rules that are specific to a given research group, institute or community. For example, one could set a constraint for the maximum light intensity based on the strongest light sources available within a given institute, thus automatically identifying light intensities that are probably wrong by an order of magnitude. Each validation message contains a severity code (Error, Warning or Info), a text message and an object which is related to the message. The view can be filtered by selecting the different levels of the severity code.

Exporting metadata

The table tab can be used to provide a summary of the metadata at each of the observation points (samples) in table format. Available metadata (the various environmental variables, events, sampling strategy) are organized in a tree structure and can be selected for output. Selected cells can be copied to the clipboard and pasted into any table-based software (e.g. Excel), which can then be used to manually collect the corresponding data (e.g. the levels of the analytes measured in the samples), each line defining a sample. A script enabling the import of the collected metadata and data into the statistical analysis software R (R Core team) is available at http://xeml.mpimp-golm.mpg.de/dnn/Home/tabid/54/Default.aspx. At present, it is only usable by advanced R users, and further work will be needed to make it available for general purpose. The operations involved in compiling metadata and data, and importing them into R, can be tested using sample files provided as Supporting information (Examples.zip).

Report

The report provides an overview of the actual experiment, with a recapitulation of the ontologies that are referenced, the terms describing growth conditions that are used, the terms that are variable, the nature of the observed material and the observation schedule including an estimation of the required time.

Environment ontology (Xeml Environment Ontology)

To be acceptable to the individual user and the community, an ontology needs to strike a balance between not being so complicated that it becomes time-consuming or even impossible to enter a description of the experiment and not being so simple that it is of little or no use for organizing, cross-checking and mining experiments. We decided to develop a new environment ontology encompassing 80 terms describing the abiotic environment. It supports a precise description of a wide range of growth conditions that are typically applied under controlled environments. The ontology was initially developed in XML because this format was the most adjustable and easiest to handle for that purpose. The XML format is currently loaded by default into Xeml Interactive Designer (see above). It is nevertheless also available in OWL and OBO formats at http://www.codeplex.com/XeO.

Each term consists of a name, a definition, a hierarchical path and three different contexts in which the value of a term can be defined. For each term, the ontology specifies a recommendation level, which reflects the attention the term probably requires in a typical experiment. Thus, in the current version, we recommend that light, temperature, relative humidity (assuming that vapour pressure deficit would be calculated) or nutrients should always be taken into account, while the description of electrical conductivity, granulometry or flooding may be restricted to specific experiments.

The value of a term can be defined in the three contexts: (1) Quantity or quality (they are mutually exclusive); (2) ResponseRange; and (3) FreeText. This is a pragmatic solution that allows the users to enter information at different levels, depending on the information they have available. For example, it is difficult to quantify the availability of nutrients in soils, but it may be possible to evaluate it roughly.

Quantitative/qualitative context

This context defines the type of data (quantitative, categorical or Boolean), including a number of specifications such as units, category names and minimum, maximum and default values. In cases where several units are defined, one is specified as the default unit, and conversions between units are provided.

Response range context

It may often be useful to document the ‘response range’, i.e. the degree to which a given treatment is optimal or stressful because the environmental variable is too low or too high. A discrete scale ranging from 0 to 9 has been chosen, as this should provide a reasonable level of detail in most situations (Fig. 4). Values ranging from 0 to 3 correspond to ‘too little’ value 4 to ‘enough’, value 5 to ‘luxurious’ and values from 6 to 9 to ‘too much’. Many environmental inputs that are required for growth can adopt values across the entire response range, for example, temperature or the light intensity. For other treatments, such as ‘pollutants’ that are probably of no benefit for plants, only values ranging from 5 (no effect) to 9 (lethal) will be relevant. The default phenotype that is used to describe the response range is biomass production. If another process is used (e.g. photosynthesis or expansive growth; leaf chlorosis; inhibition of root growth), this can be specified in the FreeText context (see below).

Choice of a ResponseRange value requires a certain level of interpretation and is clearly susceptible to inaccuracy or subjective bias. The concept of ‘response range’ also implies that there are unique optimal and stressful ranges of conditions for all processes in a plant, which is likely to be incorrect; for example, optimal conditions for photosynthesis and for expansive growth are completely different.

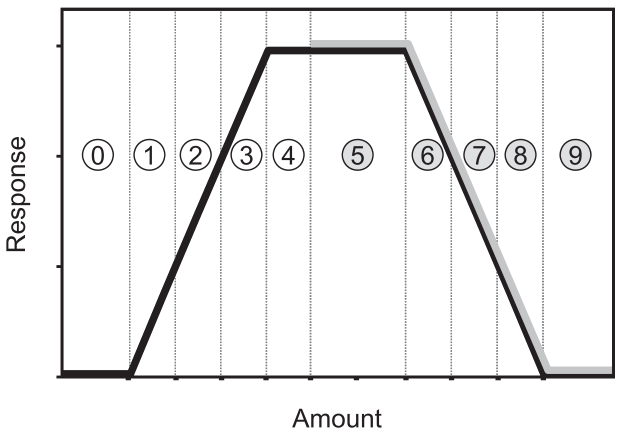


Figure 4. ResponseRange context. The black line represents the response to an essential parameter such as a nutrient. The grey line represents the response to a 'pollutant', i.e. an environmental condition that is only becoming limiting when present in excess.

Furthermore, many environmental variables may act synergistically or antagonistically on plant's performance. Despite these caveats, we have included this option for pragmatic reasons. It allows users to provide semi-quantitative information about what they think a given treatment is doing to the plant. Even though, by definition, it should be possible to derive the ResponseRange from the quantitative and qualitative metadata, this would imply an expert system which is able to infer the status of the plant from the metadata about the growth regime. As this is currently not possible, the ResponseRange provides a useful stop gap solution that harnesses the knowledge of the experimenter. Comparison of the actual treatment and the estimated response range will allow a cross-check by expert third parties. A 'rating' of what the treatments are 'thought to be' also provides a short description of the aims and scope of an experiment, and facilitates future access to the relevant data. For example, experiments in which attempts were made to apply various degrees of nitrogen limitation would be tagged and thus, easy to extract from large data sets. After extracting experiments that fall into the search space defined by a response range and a variable, it will then be possible to inspect other parts of the metadata to decide which experiments or treatments are worth retaining for further analysis. Interestingly, it may be possible to cross-check actual quantitative parameters versus the ResponseRange set, and build an expert system on the available classifications which might suggest that the experimenter is running into nitrogen starvation.

Free text context

This context enables the capture of keywords, short descriptions or specific comments about the variable under consideration. It will allow the recording of experiment-specific features that are not currently captured in the ontologies. On a midterm basis, inspection of the information entered in the free field spaces may provide information about ways in which ontologies need to be developed.

Posting of a Wiki to promote discussion of standards for the description and measurement of environmental conditions

While the choice of the experimental set-up and environmental conditions obviously depends on the aim of the experiment, it is vital to do this in the context of current knowledge and experience, and to carefully consider the best choices and also possible pitfalls with regard to the chosen environmental conditions. There are many interactions between environmental factors and plant growth (see Introduction for two examples). Furthermore, it is important to measure environmental conditions according to agreed standards, to routinely service and calibrate the instruments that are used to measure the environmental conditions, and to check carefully for possible spatial or temporal heterogeneity. Supporting information outlines how a number of the most important environmental conditions vary in nature and in controlled or semi-controlled environments, and how they are best controlled, measured and reported. It also discusses important factors that must be considered when each of these environmental conditions is used as a treatment.

These are questions that depend on community discussion and definition. To support this process, we have posted the supporting information at a collaborative website, i.e. a Wiki, at http://www.codeplex.com/XeO. The Wiki can also be accessed directly from the ontology browser by right clicking on the variable of interest. The supporting information can be commented on and developed further by the wider community of plant scientists, and requests and suggestions posted to the Wiki. In addition to providing a discussion and information forum, we hope to collect definitions and recommendations about the growth variables available in the Xeml Environment Ontology. In order to keep terms compatible between research groups, we request not to modify the ontology locally but, instead, to send a request to the corresponding authors, so that existing terms could be updated or new terms could be added centrally.

Example showing how a meta-analysis can identify sources of noise

To illustrate the advantage of documenting the experimental design, we have retrospectively entered over 40 experiments with *Arabidopsis thaliana* that were performed between 2002 and 2007 into Xeml Lab, and performed two data mining studies to extract information that was not apparent from the individual experiments. The treatments in the various experiments included harvest at six different times in the diurnal cycle of wild-type Col-0 growing in a 12 h light/12 h dark photoregime (Bläsing *et al.* 2005), harvest at the end of the light period and the end of the dark

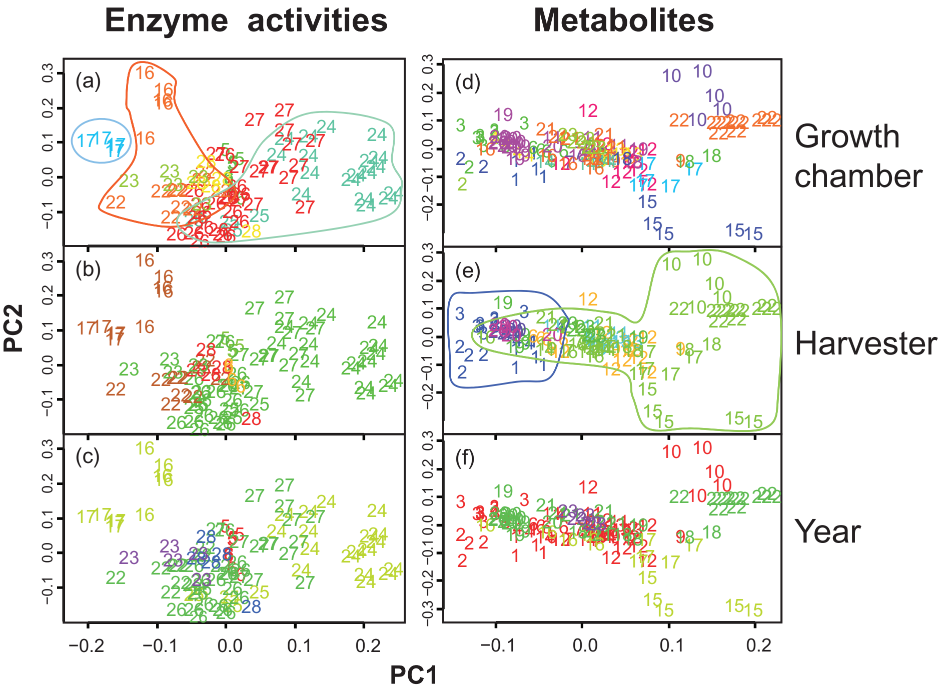


Figure 5. Principal component analysis of metadata describing the local infrastructure and sample handing of supposedly standard treatments from many separate experiments. The plots show a projection of individual samples into the first two principal components. Each experiment is coloured by the separating variable (growth cabinet, harvester, year) indicated on the left and labelled by a unique numerical identifier, to highlight differences potentially incurred by the experiment. A total of 28 separate experiments performed between 2002 and 2007 were analysed, in which there was a treatment in which wild-type *Arabidopsis* Col-0 was grown in a 12 h/12 h day/night regime, with a total light intensity between 120 and 150 μmol m^{-2} s^{-1}, a relative humidity between 80 and 90%, a temperature of 20 °C, and harvested at the end of the day. The activities of 15 enzymes and 9 metabolite levels were measured in 11 and 23 of these experiments, respectively. Examples of well-separated samples mentioned in the text are shown in (a) (growth chamber) and in (e) (harvester). Enzyme activities (a, b, c); metabolites (d, e, f); growth chamber (a, d); harvester (b, e); year (c, f).

period of wild-type Col-0 growing at different photoperiods (ranging from a 20 h light/4 h dark cycle to a 2 h light/22 h dark regime; Gibon *et al.* 2004a, 2009), different temperatures (from 20 to 8 °C; Usadel *et al.* 2008), water supply, light intensity (unpublished) and various experiments with the starchless *pgm* mutant (Gibon *et al.* 2004a; Bläsing *et al.* 2005; Usadel *et al.* 2008).

The first data mining study was performed to learn how much variation there was between supposedly identical treatments (Fig. 5). A total of 28 of the experiments included a treatment in which 5-week-old rosettes were harvested from wild-type Col-0 that was grown in supposedly standard conditions (several climate chambers from the same product line with a 12 h/12 h day/night regime, a total light intensity between 120 and 150 μmol m^{-2} s^{-1}, a relative humidity between 80 and 90%, a temperature of 20 °C) and was harvested at the end of the light period. The activities of 15 enzyme activities and the levels of nine metabolites were measured in 11 and 23 of these experiments, respectively. We noted that there was some variation between the experiments. The coefficients of variation ranged from 25 to 66% for individual enzyme activities, and from 25 to 113% for metabolites (Table 1). For four of the analytes, the coefficients of variation for this combined set of control treatments were even higher (up to two times for ferredoxin-glutamate synthase) than the coefficients of variation calculated for all the different treatments. We were intrigued by this variation between supposedly identical plant materials, and tried to identify the reason for it.

Many of the experiments were performed before we started to routinely check and collect the information about the environmental conditions that is captured in Xeml Lab. However, we were able to perform an *a posteriori* analysis using a limited set of metadata that had been collected in lab books and Excel files. This allowed us to distinguish three characteristics that could potentially affect the reproducibility of experiments: (1) the growth cabinet that was used; (2) the person who performed the harvest; and (3) the calendar year. A principal component analysis was performed with enzyme activities (11 experiments, 80 individual samples) and metabolites (23 experiments, 140 individual samples) after centring and scaling the data to avoid an excessive contribution of single enzymes (Fig. 5a–c) or metabolites (Fig. 5d–f) to the total variance. Experiments are identified by the number. Samples from a given experiment generally grouped together. The impact of the metadata variables is identified in three separate panels, in which a shared colour is used to identify samples from different experiments that (1) were grown in the same chamber Fig. 5a,d); or (2) were harvested by the same

Table 1. Comparison of the coefficient of variation (expressed in %) calculated for 24 analytes for control samples and for all samples

Analyte	Standard condition	All samples
Total amino acids	32	58
Chlorophyll *a*	29	34
Chlorophyll *b*	30	37
Fructose	70	160
Glucose-6-phosphate	49	86
Glucose	113	141
Starch	29	120
Sucrose	49	229
Protein content	25	29
Acid invertase	59	52
Alanine aminotransferase	27	36
Aspartate aminotransferase	27	30
Ferredoxin-glutamate synthase	60	32
Fructokinase	25	36
Fumarase	29	34
Glucokinase	48	59
Glucose-6-phosphate dehydrogenase	25	37
Glutamate dehydrogenase	51	46
Glycerokinase	66	70
Glutamine synthetase	30	43
NADP-isocitrate dehydrogenase	29	31
Nitrate reductase	57	62
Phosphoenolpyruvate carboxylase	29	35
Shikimate dehydrogenase	37	38

The standard condition was wild-type Col-0 that was grown in supposedly standard conditions: several climate chambers from the same product line with a 12 h/12 h day/night regime, a total light intensity between 120 and 150 μmol m^{-2} s^{-1}, a relative humidity between 80 and 90%, a temperature of 20 °C and was harvested at the end of the light period. Data were obtained from 28 experiments. Data were obtained from 23 experiments for metabolites and protein content, with a total 1642 samples of which 140 were in standard the standard condition, and from 11 experiments for enzyme activities, with a total of 336 samples including 80 in the standard condition for enzyme activities.

person (Fig. 5b,e); or (3) were grown in the same calendar year (Fig. 5c,f).

For both sets of parameters, there was little systematic impact of the calendar year (Fig. 5c,f). This shows that there had been no gradual insidious temporal drift in the plant material, growth conditions, harvesting or analytic procedures. For enzyme activities, the growth chamber had the largest systematic impact (Fig. 5a). In many cases, the outliers in PC1 are samples from a particular growth chamber, not only in one experiment (e.g. 17) but even in different experiments in which the same chamber was used (e.g. 16 and 22; 24 and 25). Other chambers consistently have a low weighting. The harvesting person had less effect (Fig. 5b). The separation in PC1 may therefore be at least partly due to the growth chambers rather than the harvester in a particular experiment. For metabolites, there was no systematic separation of samples depending on the growth chamber (Fig. 5d). Instead, samples tend to be separated that were collected by different harvesters (Fig. 5e). For example, experiments 1, 2 and 3 are outliers and were collected by the same harvester, as are many of the samples from experiments 10, 15, 17, 18 and 22 which were collected by another harvester.

Taken together, these results suggest that small discrepancies between supposedly identical growth chambers lead to variation in the observed metabolic phenotype. Unfortunately, at the time the experiments were performed, we did not use a system such as Xeml Lab, and possible fluctuations in, for example, light intensity and quality were not documented. Subsequent measurements of the light intensity and temperature after performing the present data analysis did not reveal any differences between growth cabinets. However, this merely emphasizes that such changes may be temporary, due, for example, to small differences in settings or in the quality of individual light sources. With respect to the influence of the harvester on metabolites, there is an interesting potential explanation. We have previously documented that enzyme activities are generally stable across a day and night cycle, thus integrating growth conditions over time (Gibon *et al.* 2004b, 2006). Metabolite levels may be more susceptible to how samples are quenched (Kopka *et al.* 2004). This may contribute to the increased susceptibility of the metabolite profile to the harvester. In the future, controlling and recording more growth parameters and systematically identifying the reasons for the variation in the control samples should allow us to implement procedures to decrease the variation and increase the statistical power of our experiments.

Example showing how documenting experimental conditions supports data mining to extract biological information

In a second example, data mining was performed to extract information about the response of metabolism to changes in the carbon status. This analysis used a subset of 27 experiments, including (1) harvest of wild-type Col-0 at 6 different time points in a 12 h light/12 h dark cycle (Bläsing *et al.* 2005); (2) harvest of wild-type Col-0 at the end of the light period and the end of the dark period in photoperiods ranging between a 20 h light/4 h dark and a 2 h light/22 h dark cycle (Gibon *et al.* 2004a, 2009); (3) transfer of wild-type from a 14 h light/10 h dark cycle to extended darkness for different times (Thimm *et al.* 2004; Usadel *et al.* 2008); (4) harvest of the starchless *pgm* mutant at 6 different times in a 12 h light/12 h dark cycle (Bläsing *et al.* 2005); and (5) harvest of *pgm* at the end of the light period in photoperiods ranging between a 20 h light/4 h dark and a 5 h light/17 h dark cycle; unpublished data,. As this is a retro-analysis of existing data, it is restricted to a relatively small number of traits that were analysed in most of the experiments; starch, sucrose, reducing sugars, total amino acids and total protein. The metadata files generated by XEML and the experimental data files for these experiments are provided in the supporting information.

The data files were initially searched using Xeml descriptors to identify treatments where wild-type Col-0 was

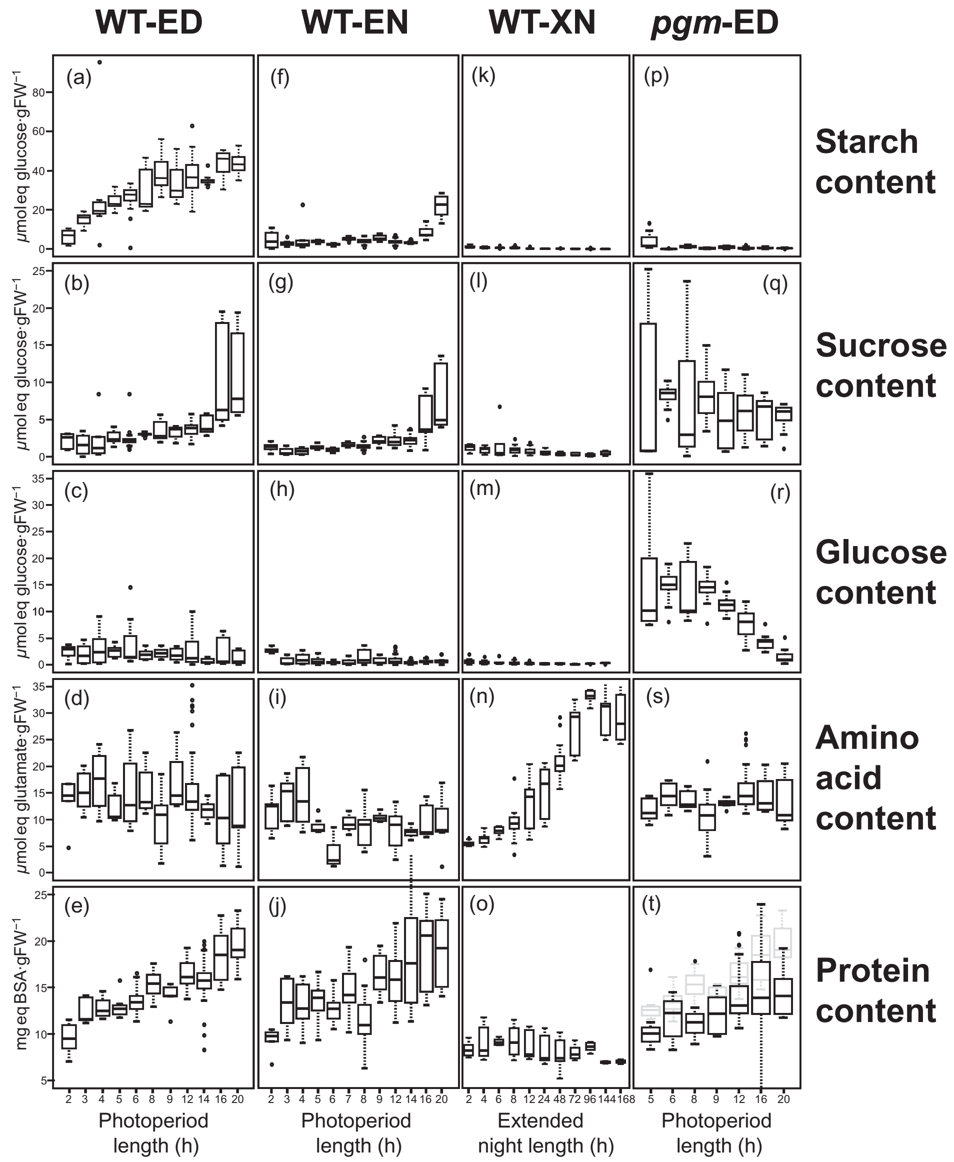


Figure 6. Box plots representing variations in starch (a, f, k, p), sucrose (b, g, l, q), glucose (c, h, m, r), total amino acids (d, i, n, s) and protein content (e, j, o, t) in rosette leaves of wild-type *Arabidopsis thaliana* grown under various day and regimes, harvested at the end of the photoperiod (a–e), at the end of the night (f–j), or following transfer into continuous darkness (k–o), and in rosette leaves of starchless *pgm* mutants grown under various day and night regimes. The light grey data shown in panel T is the corresponding data for wild-type Col-0 (extracted from panel e) to facilitate comparison with the values in *pgm*. The data were collected from a total of 23 separate experiments performed between 2002 and 2007.

grown in different photoperiods, and then split again to separate data from plants harvested at the end of the light period (WT-ED; Fig. 6a–e) and at the end of the dark period (WT-EN; Fig. 6f–j). The box plots show that shorter photoperiods led to a small decrease in starch content at the end of the light period (Fig. 6a). The starch content at the end of the night (Fig. 6f) was low but nearly constant, except in the longest photoperiod (20 h light/4 h dark). Thus, most of the starch is mobilized during the night. The average rates of starch synthesis and degradation in the light and dark period, respectively, were calculated from the difference between the amount of starch at these two time points and the length of the light and the dark period. The estimated rate of starch synthesis increased from 1.2 to 5.4 μmol hexose gFW^{-1} h^{-1} in the range of 20 to 4 h photoperiod, and then fell to 4.8 and 1.9 μmol eq glucose gFW^{-1} h^{-1} in, respectively, 3 and 2 h photoperiod. The estimated rate of starch degradation decreased from 6.3 to 0.2 μmol hexose gFW^{-1} h^{-1} in the range of 20 to 2 h photoperiod. The levels of sucrose (Fig. 6b,g) at the end of the light period were relatively constant between a 2 h and 12 h photoperiods, and rose strongly in the 16 h and 20 h photoperiod treatments. A similar picture emerged at the end of the night, except that sugar levels were about twofold lower than at the end of the light period. Amino acid levels were remarkably constant across the entire range of photoperiods (Fig. 6d,i). Thus, *Arabidopsis* adjusts to very large changes in the photoperiod and the diurnal carbon supply by altering the rates of starch synthesis and degradation while retaining relatively stable levels of sugars and amino acids. This confirms, on a much broader scale, the changes in starch allocation noted in earlier studies of photoperiod adjustment (Stitt, Bulpin & Ap Rees 1978; Chatterton & Silvius 1979, 1980, 1981; Jablonski & Geiger 1987; Mullen & Koller 1988; Lorenzen & Ewing 1992; Matt *et al.* 2001; Gibon *et al.* 2004a). However, our meta-analysis provides a further and potentially important result. The total leaf protein content (Fig. 6e,j) fell by twofold as the photoperiod decreased, falling progressively across the entire range of treatments included in this meta-analysis.

Further information was added by extracting other treatments related to changes in the carbon status. In several of the experiments, plants were darkened for up to 7 d to impose extreme carbon starvation (WT-XN; Fig. 6k–o). In these treatments, starch was negligible (Fig. 6k), sucrose and reducing sugars were low (Fig. 6l,m), but amino acid levels were as high or higher than in a light/dark cycle (Fig. 6n). Protein (Fig. 6o) decreased slowly to levels similar to those in moderately short photoperiods. The data set also included experiments in which the starchless *pgm* mutant was grown in a variety of photoperiods, ranging from 20 down to 5 h of light per 24 h cycle (*pgm*-ED; Fig. 6p–t). Starch, as expected, was always negligible (Fig. 5p). Sucrose and reducing sugars were high at the end of the light period, especially in shorter photoperiods (Fig. 6q,r, see also Caspar, Huber & Somerville 1985; Gibon *et al.* 2004a) and very low at the end of the night period (not shown, see also Caspar *et al.* 1985; Gibon *et al.* 2004a). Amino acids were similar to or slightly higher than in wild-type Col-0 in a similar photoperiod (Fig. 6s). Decreasing photoperiod length again led to a progressive decrease of total protein (Fig. 6t). Furthermore, when the genotypes are compared in the same photoperiod, *pgm* consistently contained 20–30% less protein than wild-type Col-0 (Fig. 6e,t; to aid comparison the wild-type Col-0 values are shown as pale grey in Fig. 6t, note that the very short 2, 3 and 4 h photoperiod treatments are absent in the *pgm* data set).

This meta-analysis points to a decrease in the leaf protein content as a potentially central factor in adaptation to a long term decrease in the carbon supply. Lower protein content would decrease C-utilization by decreasing the construction cost of the leaves, and by decreasing the maintenance costs that are related to protein turnover. It is well established that there is a close relation between the leaf nitrogen content and the respiration rate (James 1953). Protein turnover synthesis represents a major respiratory cost in leaves, with estimates ranging from 10 to 60% of the total leaf respiration (Penning de Vries 1975; Barneix *et al.* 1988; Bouma, Broeckhuysen & Veen 1996). Much of the respiratory substrate is directly or indirectly derived from starch breakdown. This prompted us to plot the leaf protein content against the estimated rates of starch synthesis and starch degradation in all of the photoperiod treatments (Fig. 7). There is a negative relation ($R = 0.26$; P-value $= 4.0 \times 10^{-4}$) with the rate of starch synthesis, and a positive and highly significant correlation ($R = 0.55$; P-value $= 4.4 \times 10^{-16}$) with the rate of starch degradation at night.

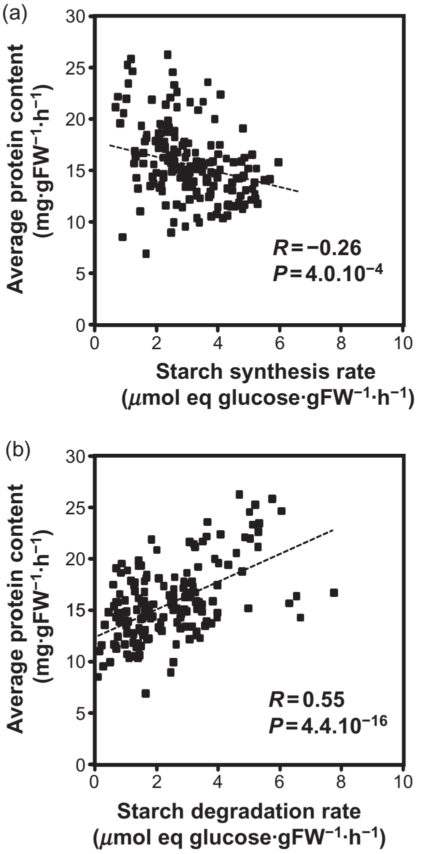


Figure 7. Starch metabolism and the protein content. Starch synthesis (a) and degradation (b) rates are plotted against the protein content measured at the end of the day. On the *x*-axis, plots represent single point calculated as the difference between the starch content measured at the end of day and the average of the starch content measured at the end of the night and within each day and night regime. The data were collected from a total of 18 separate experiments performed between 2002 and 2007 and 386 samples.

The meta-analysis provides hints about mechanisms that might underlie this dramatic change in the protein content. It is well known that growth of plants in low irradiance leads to a decreased protein content (Poorter *et al.* 2006). However, in this case, it is difficult to distinguish the effects of decreased carbon *per se* from photomorphogenetic changes in development, which alter leaf composition to optimize light absorption in different light regimes. The threefold decrease in protein content in response to changes in the photoperiod occurs independently of the momentary light intensity or quality, indicating that it may be triggered by changes in the plant carbon status. Interestingly, protein was lower in the starchless *pgm* mutant, than in wild-type Col-0 growing in the same photoperiod. Although the starchless *pgm* mutant has high levels of sugars in the light, these fall rapidly at the start of the night. Transcript profiling reveals that *pgm* experiences repeated and acute carbon starvation every night (Gibon *et al.* 2004a; Bläsing *et al.* 2005; Usadel *et al.* 2008). The enzyme activity (Gibon *et al.* 2004b) and metabolite (Gibon *et al.* 2006) profiles of *pgm* resemble carbon-starved Col-0 wild-type plants after several days in the dark. The decreased levels of protein in the *pgm* mutant may be partly due to the impact of the repeated period of carbon starvation during the night. Evidence for a general relationship between the carbon supply and leaf protein levels is provided by studies showing that tobacco plants with decreased rates of photosynthesis due to decreased expression of Calvin cycle enzymes have decreased leaf protein contents (Stitt & Schulze 1994). Of course, further mechanisms may also contribute to the lower levels of protein in short photoperiods. For example, much of the protein synthesis in leaves may occur in the light. It is well known that chloroplast protein synthesis decreases in the dark (Marin-Navarro *et al.* 2007) and it is plausible that some components of the cytosolic protein synthesis may also decrease in the dark, especially those that deliver proteins to the plastid and require plastid-encoded proteins as partners in the thylakoid complexes and ribulose 1·5-bisphosphate carboxylase/oxygenase (Rubisco). Another possibility is that protein content may be under light-dependent circadian regulation. These hypotheses, which have been extracted by meta-analysis of this data set, can be tested in future experiments.

FUTURE DEVELOPMENTS

Xeml Lab offers a convenient environment to plan and document experiments in a machine-readable manner, and provides information about the experimental design, sampling procedures and environmental conditions. In its present version, the Xeml Interactive Designer allows metadata to be captured for most experiments performed under controlled conditions. The exhaustiveness of the description will in large part rely on the relevance of the ontologies. Depending on the nature of the experiments, it might be necessary to use further extensions of the ontologies, or to add more specific ontologies. This is illustrated by the integration of existing Plant Structure Ontology into the first version of Xeml Lab.

The XEML language and its framework have been designed in such a way that experiments performed under non-controlled conditions can also be described, for example, greenhouse experiments and even field experiments, in which environmental variables cannot be predefined (or only partially). This will require that environment parameters are recorded frequently. The capture of such metadata, typically lists of time points and values, would be time-consuming in the Interactive Designer. Therefore, a dedicated tool will be developed to capture, convert and visualize such metadata automatically. This could be done by interfacing XEML with LIMS systems that store detailed information about the environmental conditions in the field of greenhouses (Köhl *et al.* 2008).

The Xeml format provides a flexible and easy way to document experiments, but each experiment still represents a single file on the hard disc and is not readily available for a workgroup or a community. Storage facilities and exploration and compilation tools will have to be developed and/or connected to make metadata stored in Xeml files and the corresponding analytic data accessible and searchable (Fig. 1). We plan a central solution consisting of a server with a database back-end and communication interfaces that support the deposition of Xeml files and perform tasks such as checking in and out, applying rules (user management, protection of documents), and searching or selecting experiments based on particular growth conditions. A major feature of XEML is that it allows metadata collected in Xeml files to be linked via providers to different types of analytic data. Thus, a tool will be implemented to explore and compile metadata and data, and to generate table outputs that can be used in analytical tools such as Excel or R. Depending on the metadata that has been recorded, it will also be possible to generate protocols that are compliant with MIAME (Minimum Information About a Microarray Experiment) or MIAMET (Minimum Information About a METabolomic experiment) standards. Such tasks will however require that all relevant data are available, and thus, that each of the relevant databases is accessible. This is currently possible at the level of a research group or of an institute, but probably not yet at the global level.

As an alternative, we will develop a zip-based package format allowing the storage, in one archive file per experiment, of the Xeml document, the version of the ontologies used and the compiled analytic data sets. Such files would then be easy to distribute via file servers, for example, as supporting information of a publication. In that context, Xeml files generated with the Xeml Interactive Designer already benefit from a global unique identifier (GUID), which allows the preservation of consistent data sets. Furthermore, such archive files could be made available via non-central solutions such as the peer-to-peer technology.

CONCLUDING REMARKS

As plant phenotypes result from interactions between genotypes and growth conditions, anything that happens to a plant during its life cycle could potentially influence the phenotype at the moment of the harvest, or the response to a given treatment. Thus, care should be taken about how plants are grown, and growth conditions should be documented as precisely and as exhaustively as possible. We propose a software-based solution that enables the simultaneous design and documentation of experiments with the help of a graphical interface and an environmental ontology. The terms included in this ontology were chosen and graded based on what we believe is important for plants growing in controlled conditions. The underlying reasons for the organization of the ontology are summarized in the supporting information; this is, however, a topic for community discussion, which we hope to promote by implementing a 'Wiki'. We think that XEML will be very useful within research groups, institute or collaborative networks. Xeml Lab was designed in such a way that documenting – even sophisticated – experiments would be as easy and as helpful as possible. Our wish is indeed that many plant scientists would adopt XEML and in doing so, would make the descriptions of their experiments compatible, because they would be talking the same language.

ACKNOWLEDGMENTS

This research was supported by the Max Planck Society, by the German Ministry for Research and Technology in the framework of the German Plant Genomics Programme Genome Analysis of the Plant Biological System (GABI) in the projects 0313110 (MapMan), 0315049A (MapMen) and 0315060E (GNADE), and by the European Community (FP7 IP AGRON-OMICS). Thanks are due to the members of the Stitt group for beta-testing and critical discussions, and in particular to M. Höhne, M. Günther and L. Bartezko for technical assistance.

REFERENCES

Atkin O.K., Loveys B.R., Atkinson L.J. & Pons T.L. (2006) Phenotypic plasticity and growth temperature: understanding interspecific variability. *Journal of Experimental Botany* **57,** 267–281.

Barneix A.J., Cooper H.D., Stulen I. & Lambers H. (1988) Metabolism and translocation of nitrogen in two *Lolium perenne* populations with contrasting rates of mature leaf respiration and yield. *Physiologia Plantarum* **72,** 631–636.

Barrett T., Troup D.B., Wilhite S.E., Ledoux P., Rudnev D., Evangelista C., Kim I.F., Soboleva A., Tomashevsky M. & Edgar R. (2007) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Research* **35,** D760–D765.

Berardini T.Z., Mundodi S., Reiser L., *et al.* (2004) Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiology* **135,** 745–755.

Bino R.J., Hall R.D., Fiehn O., *et al.* (2004) Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* **9,** 418–425.

Bläsing O.E., Gibon Y., Günther M., Höhne M., Scheible W.R., Osuna D. & Stitt M. (2005) Genomics analysis of diurnal changes of expression in *Arabidopsis* reveals major contributions from carbohydrates and circadian regulation. *The Plant Cell* **17,** 3257–3281.

Bouma T.J., Broeckhuysen A.G.M. & Veen B.W. (1996) Analysis of root respiration of *Solanum tuberosum* as related to growth, ion uptake and maintenance of biomass. *Plant Physiology and Biochemistry* **34,** 795–806.

Bradford M.M. (1976) Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding. *Analytical Biochemistry* **72,** 248–254.

Brazma A., Hingamp P., Quackenbush J., *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics* **29,** 365–371.

Caspar T., Huber S.C. & Somerville C. (1985) Alterations in growth, photosynthesis, and respiration in a starchless mutant of *Arabidopsis thaliana* (I) deficient in chloroplast phosphoglucomutase activity. *Plant Physiology* **79,** 11–17.

Chatterton N.J. & Silvius J.E. (1979) Photosynthate partitioning into starch in soybean leaves. 1. Effects of photoperiod versus photosynthetic period duration. *Plant Physiology* **64,** 749–753.

Chatterton N.J. & Silvius J.E. (1980) Photosynthate partitioning into leaf starch as affected by daily photosynthetic period duration in six species. *Physiologia Plantarum* **49,** 141–144.

Chatterton N.J. & Silvius J.E. (1981) Photosynthate partitioning into starch in soybean leaves. 2. Irradiance level and daily photosynthetic period duration effects. *Plant Physiology* **67,** 257–260.

Dennis E.S. & Peacock W.J. (2007) Epigenetic regulation of flowering. *Current Opinion in Plant Biology* **10,** 520–527.

Edgar R. & Barrett T. (2006) NCBI GEO standards and services for microarray data. *Nature Biotechnology* **24,** 1471–1472.

Frawley W.J., Piatetsky-Shapiro G. & Matheus C.J. (1991) Knowledge discovery in databases: an overview. In *Knowledge Discovery in Databases* (eds G. Piatetsky-Shapiro & W.J. Frawley) pp. 1–27. AAAI Press/MIT Press, Menlo Park, CA/Cambridge, MA, USA.

Gibon Y., Bläsing O.E., Palacios-Rojas N., Pankovic D., Hendriks J.H.M., Fisahn J., Höhne M., Günther M. & Stitt M. (2004a) Adjustment of diurnal starch turnover to short days: depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period. *The Plant Journal* **39,** 847–862.

Gibon Y., Bläsing O.E., Hannemann J., Carillo P., Höhne M., Hendriks J.H.M., Palacios N., Cross J., Selbig J. & Stitt M. (2004b) A robot-based platform to measure multiple enzyme activities in *Arabidopsis* using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *The Plant Cell* **16,** 3304–3325.

Gibon Y., Usadel B., Blaesing O.E., Kamlage B., Höhne M., Trethewey R. & Stitt M. (2006) Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. *Genome Biology* **7,** R76.

Gibon Y., Pyl E.-T., Sulpice R., Lunn J.E., Höhne M., Günther M. & Stitt M. (2009) Adjustment of growth, starch turnover, protein content and central metabolism to a decrease of the carbon supply when *Arabidopsis* is grown in very short photoperiods. *Plant, Cell & Environment*. doi: 10.1111/j.1365-3040.2009.01965.x

Granier C. & Tardieu F. (1998) Is thermal time adequate for expressing the effects of temperature on sunflower leaf development? *Plant, Cell & Environment* **21,** 695–703.

Gruber T.R. (1993) Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies* **43,** 907–928.

Hastings I.M., Lalloo D.G. & Khoo S.H. (2006) The daunting process of MIAME. *Nature* **444,** 31.

Hendriks J.H.M., Kolbe A., Gibon Y., Stitt M. & Geigenberger P. (2003) ADP-glucose pyrophosphorylase is activated by post-translational redox-modification in response to light and to sugars in leaves of *Arabidopsis* and other plant species. *Plant Physiology* **133,** 838–849.

Ilic K., Kellogg E.A., Jaiswal P., *et al.* (2007) The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiology* **143,** 587–599.

Ingestad D. (1982) Relative addition rate and external concentration – driving variables used in plant nutrition research. *Plant, Cell & Environment* **5,** 443–453.

Jablonski L.M. & Geiger D.R. (1987) Responses of sugar-beet plant morphology and carbon distribution to shortened days. *Plant Physiology & Biochemistry* **25,** 787–796.

Jaiswal P., Avraham S., Ilic K., *et al.* (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and Functional Genomics* **6,** 388–397.

James W.O. (1953) The terminal oxidases in the respiration of the embryos and young roots of barley. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **141,** 289–299.

Jelitto T., Sonnewald U., Willmitzer L., Hajirezeai M. & Stitt M. (1992) Inorganic pyrophosphate content and metabolites in potato and tobacco plants expressing Escherichia-coli pyrophosphatase in their cytosol. *Planta* **188,** 238–244.

Köhl K.I., Basler G., Lüdemann A., Selbig J. & Walther D. (2008) A plant resource and experiment management system based on the Golm plant database as a basic tool for omics research. *Plant Methods* **4,** 11.

Kopka J., Fernie A.R., Weckwerth W., Gibon Y. & Stitt M. (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biology* **5,** R109.

Kotak S., Larkindale J., Lee U., von Koskull-Doring P., Vierling E. & Scharf K.D. (2007) Complexity of the heat stress response in plants. *Current Opinion in Plant Biology* **10,** 310–316.

Krizek D.T. (1982) Guidelines for measuring and reporting environmental conditions in controlled-environment studies. *Physiologia Plantarum* **56,** 231–235.

Langhans R.W. & Tibbitts T.W. (eds) (1997) *Plant Growth Chamber Handbook*. North Central Regional Research Publication No. 340, Iowa State Agriculture and Home Economics Experiment Station Report No. 99. Iowa State University, Ames, IA, USA.

Lorenzen J.H. & Ewing E.E. (1992) Starch accumulation in leaves of potato (*Solanum tuberosum* L.) during the first 18 days of photoperiod treatment. *Annals of Botany* **69,** 481–485.

McLaren C.G., Bruskiewich R.M., Portugal A.M. & Cosico A.B. (2005) The international rice information system. A platform for meta-analysis of rice crop data. *Plant Physiology* **139,** 637–642.

Marin-Navarro J., Manuell A.L., Wu J. & Mayfield S.P. (2007) Chloroplast translation regulation. *Photosynthesis Research* **94,** 359–374.

Matt P., Geiger M., Walch-Liu P., Engels C., Krapp A. & Stitt M. (2001) Elevated carbon dioxide increases nitrate uptake and nitrate reductase activity when tobacco is growing on nitrate, but increases ammonium uptake and inhibits nitrate reductase activity when tobacco is growing on ammonium nitrate. *Plant, Cell & Environment* **24,** 1119–1137.

Mullen J.A. & Koller R. (1988) Daytime and nighttime carbon balance and assimilate export in soybean leaves at different photon flux densities. *Plant Physiology* **86,** 880–884.

Pearson K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2,** 559–572.

Penning de Vries F.W.T. (1975) The cost of maintenance processes in plant cells. *Annals of Botany* **39,** 77–92.

Poorter H., Pepin S., Rijkers T., De Jong Y., Evans J.R. & Körner C. (2006) Construction costs, chemical composition and payback time of high- and low-irradiance leaves. *Journal of Experimental Botany* **57,** 355–371.

R Development Core Team (2006) *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Austria. ISBN 3-900051-07-0. URL http://www.R-project.org

Rayner T.F., Rocca-Serra P., Spellman P.T., *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* **7,** 489.

Sadok W., Naudin P., Boussuge B., Muller B., Welcker C. & Tardieu F. (2007) Leaf growth rate per unit thermal time follows QTL-dependent daily patterns in hundreds of maize lines under naturally fluctuating conditions. *Plant, Cell & Environment* **30,** 135–146.

Stitt M. & Schulze D. (1994) Does Rubisco control the rate of photosynthesis and plant growth – an exercise in molecular ecophysiology. *Plant, Cell & Environment* **17,** 465–487.

Stitt M., Bulpin P.V. & Ap Rees T.A. (1978) Pathway of starch breakdown in photosynthetic tissues of *Pisum sativum*. *Biochimica et Biophysica Acta* **544,** 200–214.

Stoeckert C.J., Causton H.C. & Ball C.A. (2002) Microarray databases: standards and ontologies. *Nature Genetics* **32,** 469–473.

Thimm O., Bläsing O., Gibon Y., Nagel A., Meyer S., Kruger P., Selbig J., Muller L.A., Rhee S.Y. & Stitt M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* **37,** 914–939.

Thimm O., Bläsing O.E., Usadel B. & Gibon Y. (2006) Evaluation of the transcriptome and genome to inform the study of metabolic control. In *Control of Primary Metabolism in Plants* (eds B. Plaxton & M. McManus), Annual Plant Reviews **22**, 1–23. Blackwell Publishing, Oxford, UK.

Usadel B., Bläsing O.E., Gibon Y., Retzlaff K., Höhne M., Günther M. & Stitt M. (2008) Global transcript levels respond to small changes of the carbon status during progressive exhaustion of carbohydrates in *Arabidopsis* rosettes. *Plant Physiology* **146,** 1834–1861.

Zimmermann P., Schildknecht B., Craigon D., *et al.* (2005) MIAME/Plant – adding value to plant microarrray experiments. *Plant Methods* **2,** 1.

Received 25 November 2008; received in revised form 28 January 2009; accepted for publication 29 January 2009

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Supporting Information. Xeml Interactive Designer can be downloaded and installed at http://xeml.mpimp-golm.mpg.de/dnn/Resources/tabid/56/Default.aspx. The three libraries composing the Xeml Framework can also be downloaded separately from this website, in case they are needed for custom programming. A script enabling the import of compiled files in csv format (xeml metadata and data) into the statistical analysis software package R, and to retrieve the time of the day at which samples have been taken, can also be downloaded at the address.

Appendix S1. Environmental conditions. This supporting information outlines how a number of the most important environmental conditions vary in nature, and how they are best measured, controlled and reported. Furthermore, we discuss briefly some aspects to consider when these environmental conditions are used as a treatment and provide a list with – for each treatment – some of the most consistent changes in plants.

Supporting Examples. Data (csv files) and metadata (xeml files) used in meta-analysis shown in Figure 6 are provided in a compressed archive (Examples.zip).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.