

Taxonomy-Based Glyph Design—with a Case Study on Visualizing Workflows of Biological Experiments

Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone, Jim Davies, and Min Chen

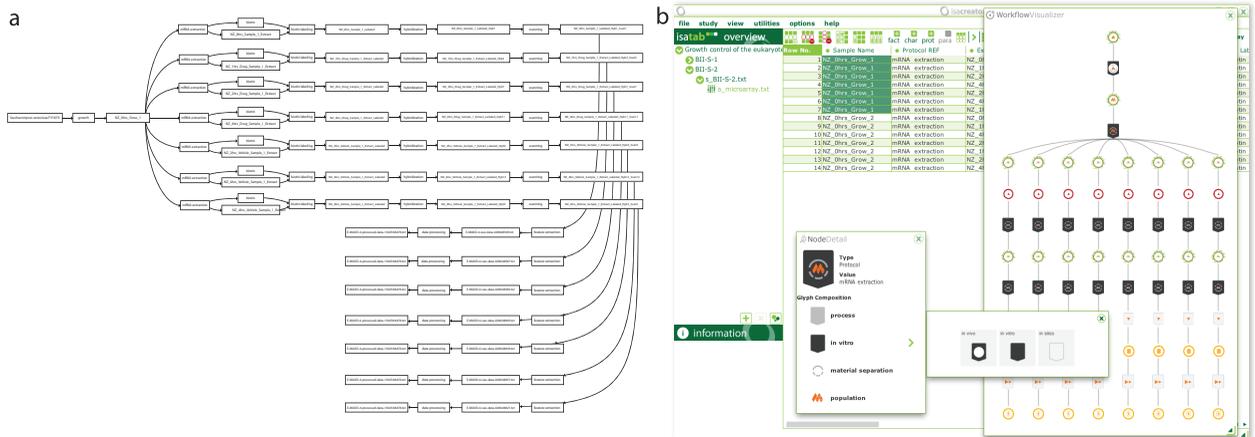


Fig. 1. a) Workflow as rendered currently using toolkits such as GraphViz. b) We propose to replace the textual labels with glyphs, while allowing interactive access to detailed descriptions. This makes it easy to gain an overview, search components and compare workflows. The screenshot shows a prototype developed within ISAcceptor, a system for capturing biological experiment metadata.

Abstract—Glyph-based visualization can offer elegant and concise presentation of multivariate information while enhancing speed and ease in visual search experienced by users. As with icon designs, glyphs are usually created based on the designers' experience and intuition, often in a spontaneous manner. Such a process does not scale well with the requirements of applications where a large number of concepts are to be encoded using glyphs. To alleviate such limitations, we propose a new systematic process for glyph design by exploring the parallel between the hierarchy of concept categorization and the ordering of discriminative capacity of visual channels. We examine the feasibility of this approach in an application where there is a pressing need for an efficient and effective means to visualize workflows of biological experiments. By processing thousands of workflow records in a public archive of biological experiments, we demonstrate that a cost-effective glyph design can be obtained by following a process of formulating a taxonomy with the aid of computation, identifying visual channels hierarchically, and defining application-specific abstraction and metaphors.

Index Terms—Glyph-based techniques, taxonomies, design methodologies, bioinformatics visualization.

1 INTRODUCTION

Glyph-based visualization is a class of visual representations where a collection of small visual objects (referred to as *glyphs*) are used to encode different attribute dimensions of an input data space. A good glyph design can enable users to conduct visual search more efficiently during interactive visualization, and facilitate effective learning, memorizing and using the visual encoding scheme. A less effective visual design may suffer from various shortcomings such as being perceptually confusing, semantically ambiguous, difficult to learn and remember, or unable to accommodate low-resolution display devices.

Most accepted designs have undergone an enduring process of evolution, refinement and standardization. One could not and should not remove the necessity for such processes in glyph design. Meanwhile, the design process for a glyph set usually relies on an assortment of creativity, artistic skill, domain knowledge, intuition about users, and sometimes personal preference. While most of these qualities are absolutely helpful and some are inevitable, it is highly desirable to introduce “systematicism” and objectivity into the design process.

To demonstrate our approach, we elected the field of experimental biology as a test bed for developing a systematic methodology to create a glyph-based visual mapping for visualizing experimental design and experimental processes. While experimental design is at the heart of biological data evidence gathering, representation of such information is confined to either verbose description or ineffective representations. This claim is backed by a survey of public biodata repositories, devised to ensure data perennity, scientific scrutiny and reproducibility. As illustrated in Fig. 1(a), workflows are traditionally drawn as text-based diagrams. Text labels are indices of concepts, and usually they do not encode multivariate information directly. Text-labeled boxes are costly in terms of display space usage as well as time required for visual search since parsing and interpretation of text is a slow post-attentive process [59]. They are particularly ineffective when one needs to compare between different workflows or to identify unusual or missing components in a workflow.

- Eamonn Maguire is with Oxford e-Research Centre and Department of Computer Science, University of Oxford, UK. E-mail: eamonn.maguire@st-annes.ox.ac.uk.
- Philippe Rocca-Serra, Susanna-Assunta Sansone and Min Chen are with Oxford e-Research Centre, University of Oxford, UK. E-mail: {philippe.rocca-serra,susanna-assunta.sansone,min.chen}@oerc.ox.ac.uk.
- Jim Davies is with Department of Computer Science, University of Oxford, UK. E-mail: jim.davies@cs.ox.ac.uk.

Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.

For information on obtaining reprints of this article, please send e-mail to: ivcg@computer.org.

Therefore, exploring an alternative in the form of glyph based rep-

representations, which are at the core of many successful schematic diagrams in the history of sciences, engineering and business management, offers a potentially more effective means for depicting these workflows. This presents us with an interesting case study where other types of design approaches would not be appropriate, especially in dealing with several hundreds of conceptual names.

It is important to note that domain experts are in general more willing and able to learn and memorize an encoding scheme in order to improve the accuracy and efficiency of routine tasks. At the same time, it is also necessary for an encoding scheme to facilitate effective learning and remembering through appropriate abstraction and metaphors.

Motivated by this case study, we made an observation that when a large number of concepts (or taxa) are organized into a taxonomic tree, the hierarchy typically represents an ordering of different categorization schemes. The schemes that are higher-up in the taxonomy (closer to the root) are usually considered to be more important, which reflect their conceptual coverage, frequency of usage, domain-specific convention, and some other measurable factors. In terms of visual encoding, higher up schemes should ideally be mapped to visual channels that have more discriminative capacity. Hence, we can explore the parallel between the taxonomic hierarchy and the ordering of visual channels based on discriminative capacity to formulate a systematic and relatively objective process for designing a glyph set. This approach addresses one of the most common criticisms of data glyphs: the data to visual attribute mapping bias [58]. Given a large data repository that encompasses many concepts, our design process is composed of four major steps: (i) gathering and processing raw metadata for obtaining a set of taxa (names); (ii) formulating a taxonomy based on a set of categorization schemes (Section 4); (iii) carrying out visual design, which includes the sub-process of determining the ordering of visual channels, proposing optional visual mappings, and identifying metaphoric abstractions and associations (Section 5); and (iv) implementing a glyph-based visualization system, in our case, for depicting workflows of biological experiments (Section 6).

Similar to most design processes, it is helpful to conduct all stages in a progressive and iterative manner. As glyph-based visualization is normally deployed in a specific application domain, it is important to involve domain experts at every stage of the design process. During this work, we met with domain specialists on a weekly basis.

2 RELATED WORK

In this section, we give a brief overview of two most relevant areas in visualization, *glyph-based visualization* and *workflow visualization*. The remainder background information includes the biological data management, perceptual guidance, and categorization algorithms, which will be described in the following sections where the relevant technical details are discussed.

2.1 Glyph-based Visualization

There are many examples of glyph usage in the literature spanning many disciplines, especially in conjunction with many schematic diagrams. Bertin considered a range of simple glyphs in the context of geo-information visualization [7]. Ward surveyed the use of glyphs in visualization, and discussed a number of bias issues, design approaches, and layout options [58]. Ropinski *et al.* conducted a survey on glyph-based techniques in medical visualizations [48]. In visualization, a number of interesting glyph designs have been presented with noticeable impact on a large range of applications (e.g., medicine [25], software [13], text [47], and scientific computation [62]). Furthermore, Ribarsky *et al.* developed an editing system, Glyphmaker [45], to assist the process of designing glyphs. Post *et al.* proposed a language, Icon Modeling Language, for creating glyphs and glyph-style contours [42].

There have been some recent efforts to use glyph-based visualization in biological sciences. One noticeable attempt is the *Systems Biology Graphical Notation* (SBGN) [28]. The design of SBGN makes use of the notional representations of UML with some modifications to describe the biological entities and their interactions within a biological system. GenoCAD [11] provides a grammar-based language for

representing and searching DNA sequences and for building genetic constructs from DNA sequences, facilitating the use of conventional link diagrams. Both systems rely heavily on text labels. In handling a large collection of workflows, they exhibit a few shortcomings, such as inefficiency in using display space and ineffectiveness in supporting some visualization tasks, such as comparisons and novelty or anomalies detection in workflows.

In the literature, several authors encouraged glyph designers to consider visual perception when constructing glyphs [58, 23]. Others examined the design space of icons, which normally encode less information than glyphs (e.g., [21, 29]). There were perceptual studies showing the merits of icons over text labels (e.g., [34, 41]), as well as those showing the contrary (e.g., [60]). Building on such discussions, this work aims to address a methodological need for a systematic approach to glyph design for applications where large collections of concepts need to be visually encoded using glyphs.

2.2 Workflow Visualization

The need for *workflow visualization* is pervasive across many different domains. For example, in business and management the Gantt chart is a form of text based workflow visualization, while BPMN (Business Process Model and Notation) and EPC (Event-driven Process Chain) make use of icons to enrich text labels. In engineering disciplines, various schematic diagrams, such as UML (Unified Modeling Language), Petri-net and circuit diagrams, are used to convey data flow and process interaction.

In this work, we consider workflows used to describe biological experiments. This class of workflow visualization renders the processes enacted on biological materials in an experiment (e.g., removal of blood sample from patient). There has been little effort to develop tools that address the domain-specific needs. Scientists usually use generic text-based graph drawing tools such as GraphViz [3].

One related aspect is *pathway visualization*, which is concerned with the rendering of cellular biological processes. An array of pathway visualization tools have been developed, some of which incorporate glyphs. For example, VANTED [22] overlays glyphs representing gene expression, enzyme and metabolite profile data on top of pathway diagrams from KEGG [37]. GENeVis [9], which also employs glyphs to represent multivariate data, is a comprehensive visualization toolkit for exploration of pathways in conjunction with temporal data. GenoCAD has at its core a workflow generation system and relies on their glyph library for rendering of the different biological processes within a cell. SBGN-ED [14], is an add-on for the VANTED software and performs pathway creation using the aforementioned SBGN [28] visual language.

3 MOTIVATION AND OVERVIEW

The advent of massively parallel techniques such as DNA microarray, mass-spectrometry based proteomics and metabolomics and next generation sequencing enables molecular biologists to collect, process and manipulate biological signals on a new scale. Big data in biology is a reality. There has been serious effort in biology to ensure quality of experimental records by providing data archiving infrastructure and defining standards for adequate annotation and sufficient details for recapitulation of results. A number of molecular biology signature databases have been or are being established to cover the main molecular dimensions: transcript, protein and metabolites. The captured metadata mostly revolves around a similar configuration where sets of samples corresponding to different conditions are processed, measurements produced and analysis performed, delivering data that needs to be handled and interpreted. While the availability of such metadata facilitates comparison across datasets and enables meta-analysis, providing means to serve an overview of experimental design can assist analysts and data managers alike, from data selection according to relevancy, to error detection such as imbalances, irregularities or inconsistencies in records.

Task Analysis. In the context of meta-databases for experimental records, there are two main groups of users. The majority of users are those who create data for their biological experiments or retrieve data

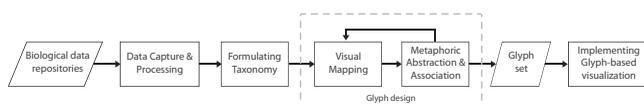


Fig. 2. A systematic process for creating glyph based representations.

relevant to their scientific interests. Their tasks include: (a) entering and editing their own experimental records; (b) transforming workflow records to schematic diagrams for comparison, external memorization, publication and education; (c) uploading and downloading datasets; (d) querying and searching for relevant experiments; (e) understanding workflows of existing experimental designs; (f) identifying similarity and difference between workflows for a common task; (g) understand pooling events and sample relations (derivation).

The second group of users are curators who manage data archives. As biology or bioinformatics scientists themselves, they perform the above-mentioned tasks (d)-(g) frequently, and (b)-(c) occasionally. In addition, they also carry tasks for: (h) checking syntactic correctness of submitted workflows; (i) checking semantic correctness of submitted workflows which usually involves reading the associated publications then comparing the submitted workflows with those described in the publications; (j) interacting with authors of submitted workflows to clarify any inconsistency and misunderstanding; (k) making appropriate corrections in submitted records where necessary; (l) augmenting with appropriate annotation based on additional information found in the associated publications; (m) providing authors with submission feedback. (n) forming an up-to-date overview about the experiments in the data archives, and maintaining an insight about the provenance of the archives; (o) analyzing the grouping, replication, distribution and trend of experiments; (p) analyzing ambiguities, errors and uncertainty in experimental recording, and identifying the need to enrich and refine the relevant standards for ontology, labelling and annotation. (p) providing tools to assist users in creating meta-data from raw data.

It is not difficult to observe that effective workflow visualization can significantly improve users' capability in performing tasks b, e, f, g, h, i, j, l, n, o and p.

While it is necessary to evolve a glyph-based design for workflow visualization over a period, it is also important not to make the first step in an ad hoc manner. Such a process does not scale well with the requirements of the application concerned where a large number of concepts are to be encoded using glyphs. We thus adopt a new systematic process for glyph design by exploring the parallel between the hierarchy of concept categorization and the ordering of discriminative capacity of visual channels. Fig. 2 depicts this process.

Data Capture and Processing. We first retrieved workflow metadata from a biological repository (content of the ArrayExpress archive), through use of a multi-threaded harvesting operation. All experimental workflows were then converted using a MAGE-Tab to ISA-Tab converter, the latter format being more general than the former and can be used to carry metadata payloads for many types of experiments, making the data processing program more "future-proof".

From the 21,000 ISA-Tab files, we extracted all names of protocols (processes) and biological materials, chemical materials, devices and data used in annotation. We also computed the occurrence of each material and process found in the entire set of ISA-Tab files, which have been used as one of the metrics in the next stage (see Section 5 for details). This step results in 61 process names and 3492 names of inputs and outputs (e.g., biological and chemical materials, device measurements and data).

Taxonomy Formulation. Since a taxonomy for such a large collection of terms is absent, we, for starters, established a set of categorization schemes with the help of domain experts. For example, one of the schemes for categorizing processes may be based on different biological inputs to a process (e.g., molecule, cell, organism, etc.). Another scheme may be based on processing methods (e.g., perturbation, combination, etc.). We computed the quality measures of each scheme based a set of generic metrics (see Section 4) then cre-

ated a taxonomic tree by recursively selecting the best categorization scheme based on the quality measures. We finalized the organization of the taxonomy by allowing domain experts (2 co-authors) to make adjustments according to domain-specific conventions. All leaf nodes of the resulting taxonomy tree are names extracted from the workflow metadata. All non-leaf nodes represent a categorization scheme.

Glyph Design. First, we established an ordering of commonly-used visual channels based on the literature focused on perception and visualization (see Section 5). This allows us to systematically compare the order of categorization schemes in the taxonomy with the order of different visual channels. Ideally, schemes at the upper level of the tree can be mapped to visual channels which are more prominent to our visual system.

We then proceeded to the glyph design process involving two intertwined sub-processes for **visual mapping** and **metaphoric abstraction and association**. We proposed various options of visual channels for each categorization scheme featured in the taxonomic tree. We considered the merits of these visual channels and identify potential conflicts with the visual channels that have been assigned to other schemes. With the direct help from domain experts, we tried to identify a metaphoric abstraction or association for each design option proposed. We evaluated and recorded the intuitiveness and suitability of the abstraction and metaphor association.

Informed by the results the two sub-processes, we finalized a glyph set by selecting a design option for each scheme, while maintaining the order of discriminative capacity of visual channels, minimizing conflicts between different channels, and maximizing the use of metaphoric abstraction and association.

Implementation of Glyph-based Visualization. We then integrated the glyph set with a layout algorithm to form a prototype system for workflow visualization (see Section 6). For demonstration purposes, we tested the workflow visualization in conjunction with ISAcceptor, a popular domain agnostic biological experiment metadata capture tool.

4 TAXONOMY GENERATION

Given a large collection of concepts, we should ideally make use of a standard taxonomy, where non-leaf nodes represent different categorization schemes and each scheme provides subclasses that lead to different sub-trees. In many circumstances, however, there is no agreed taxonomic tree as establishing a standard categorization requires non-trivial scientific effort that often spans over a few decades.

The algorithm described below is not intended to establish a semantic-rich taxonomic tree for classifying concepts. It is designed purely for addressing the needs for ordering various categorization schemes (when there is no such an agreed order) to aid glyph design, and for devising a useful tree data structure in implementing the mapping between concepts and glyphs in workflow visualization.

Hence, the criteria used for structuring the tree are based on usage of the concepts and the structural quality of the tree to be constructed.

Taxonomy is a long standing concept that can be traced back to 3000BC [32]. Automatic taxonomy generation has been an active field in computer science and computational biology with existing work largely focusing on clustering algorithms (e.g., [27]). Many such algorithms assume the availability of similarity measures for ordering entities rather than relying on the existence of individual categorization schemes. In these algorithms, there is usually no attempt towards application of a metric for creating meaningful non-leaf nodes in the resulting tree. In this work, it is essential to keep each categorization scheme as a non-leaf node unless it is redundant.

4.1 Ordering Classification Schemes

Let $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ be a set of concepts to be classified. In our application, this is the set of all valid names of biological processes and IOs (inputs and outputs) in the database. Each concept x_i is associated with a scalar value, $\mu_i \in [0, 1]$, indicating its frequency of usage in relation to the total occurrence of all concepts in the database. There are a number of categorization schemes, $\mathcal{S} = \{S_1, S_2, \dots, S_m\}$. Each scheme, S_k divides concepts into several classes, $c_1^k, c_2^k, \dots, c_{l_k}^k$. The

relationship between the concept set \mathcal{X} and different classification schemes can thus be represented by a Boolean matrix, \mathbf{A} , where each element $\alpha[i, j, k] = 1$ if concept x_i belongs to the j^{th} class of scheme S_k ; otherwise $\alpha[i, j, k] = 0$. In the context of feature-based categorization, one can also view each scheme S_k as a feature, and each class under S_k as a particular type of this feature. Without losing generality, we assume that the classes under the same S_k are disjoint. It is also possible that a concept does not possess the k^{th} feature, and hence does not belong to any class under S_k .

Given the above categorical information about the concept set \mathcal{X} , one can choose a categorization scheme $S_k \in \mathcal{S}$, which will divide \mathcal{X} into a number of disjoint subsets corresponding to classes $c_1^k, c_2^k, \dots, c_{l_k}^k$ and c_0^k . The subset c_0^k contains those concepts which S_k is unable to classify. The partitioning process can be repeated recursively by applying one of the remaining categorization schemes in \mathcal{S} to each subset. This results in a hierarchical categorization tree, which defines a taxonomy for the concepts set \mathcal{X} .

The ordering of the schemes in \mathcal{S} thus determines the taxonomic structure of \mathcal{X} . It is not difficult to anticipate that many criteria can be used to determine the ordering for a given concept set. Some criteria will no doubt encode application-specific semantics, and some may be subjective or debatable. However, there are also some common-sense criteria that are generic to most applications. These include:

Coverage. The number of concepts that can be classified by scheme S_k is a capacity measure of S_k . The more concepts that S_k can classify (i.e., the fewer in c_0^k), the better. The measure, which is normalized by the set size $|\mathcal{X}| = n$, can be defined as:

$$M_1(S_k) = \frac{\sum_{i=1}^n \max_{1 \leq j \leq l_k} (\alpha_{i,j,k})}{n}. \quad (1)$$

Potential Usage. A categorization scheme that is higher up in the taxonomical tree is expected to be used more often in the application concerned. The occurrence frequencies of concepts, μ_i , enable us to estimate the potential usage of a classification scheme as follows:

$$M_2(S_k) = \frac{\sum_{i=1}^n \mu_i \max_{1 \leq j \leq l_k} (\alpha_{i,j,k})}{\sum_{i=1}^n \mu_i}. \quad (2)$$

Subtree Balance. Having a balanced node distribution in a tree is a desirable property of a tree structure. It prevents a tree from having an excessive height, which corresponds to the need for more visual channels. Let l_k denote the number of subclasses in categorization scheme S_k , τ_j be the number of concepts in each subclass c_j^k , $j = 1, 2, \dots, l_k$ and σ_τ and $\bar{\tau}$ be the standard deviation and mean respectively of $\tau_1, \dots, \tau_{l_k}$. We can measure the level of balance as follows:

$$M_3(S_k) = \begin{cases} 0 & \sum_{i=1}^{l_k} \tau_i = 0 \\ 1 & \sigma_\tau < \varepsilon (\varepsilon > 0) \\ P(\bar{\tau} \pm 1 | N_{\bar{\tau}, \sigma_\tau}) & \bar{\tau} > 0 \ \& \ \sigma_\tau > \varepsilon \end{cases} \quad (3)$$

where P is the probability that a value within $[\bar{\tau} - 1, \bar{\tau} + 1]$ falls under the curve given by the normal distribution N with $(\bar{\tau}, \sigma_\tau)$ [40]. There are two special cases. When all values of τ_j are 0, S_k cannot classify any concept; hence the metric returns a zero score. When $\sigma_\tau = 0$, the subtree is totally balanced; hence the metric returns one. As σ_τ approaches zero, the function P becomes numerically unstable, we use a cut-off value ε to prevent this. In this work, we set $\varepsilon = 0.00001$. The normal distribution is preferred over a χ -test, as χ may not be reliable when l_k is a small number.

Number of Subclasses. All visual channels used in glyphs have limited discriminative capacity. A higher number of subclasses in a scheme would require visual encoding to have more codewords (e.g., more colors or more shape types), which will increase users' cognitive load in learning, remembering, and recognizing the codewords. It is thus more desirable to have a smaller number of subclasses, except that a categorization scheme with fewer than 2 sub-classes is useless. Let η^\top be an up-limit for the number of codewords, which is set to

10 in this work. We introduce the following metric to measure the discriminative capacity of a scheme:

$$M_4(S_k) = \begin{cases} 0 & \eta_k < 2 \\ \frac{\eta^\top - \eta_k + 2}{\eta^\top} & 2 \leq \eta_k \leq \eta^\top \\ \frac{1}{\eta^\top} & \eta_k > \eta^\top \end{cases} \quad (4)$$

Although the above four metrics have been normalized to ensure their functional values within the $[0, 1]$ domain, the distribution of the values for different schemes can still be rather application-specific and may vary substantially between different metrics. This may lead to inconsistency in combining these metrics.

We thus provide an optional linearization filter for these metrics by mapping the values obtained using a each metric M_i into fractional ranking numbers, which are then normalized into the $[0, 1]$ domain with 1 being the best and 0 the worst. For example, consider a set of six schemes with metric values:

$$(S_1, 0.5), (S_2, 0.3), (S_3, 0.54), (S_4, 0.8), (S_5, 0.85), (S_6, 0.6).$$

We first sort the set:

$$(S_2, 0.3), (S_1, 0.5), (S_3, 0.54), (S_6, 0.6), (S_4, 0.8), (S_5, 0.85)$$

We then obtain the normalized ranking values via the *distance for ordinal variables* function $\sigma = \frac{r-1}{R-1}$ where R is the top rank and r is the rank position for each schema.

$$(S_2, 0), (S_1, \frac{1}{5}), (S_3, \frac{2}{5}), (S_6, \frac{3}{5}), (S_4, \frac{4}{5}), (S_5, \frac{5}{5}).$$

We denote this filter as a function $R(S_k, M_i, \mathcal{S})$. Using the above set of metrics in conjunction with the filter R , we can derive a combined metric as

$$M(S_k) = \frac{\sum_1^4 \omega_t R(S_k, M_t(S_k), \mathcal{S})}{\sum_1^4 \omega_t}.$$

where $\omega_t, t = 1, 2, 3, 4$ are user-adjustable weights for the four individual metrics. Similar to weights in clustering algorithms, these weights need to be used with care as they introduce additional semantics into the ordering algorithm.

Equipped with the combined metric M , the algorithm for establishing an order of different schemes in \mathcal{S} can be described as follows:

```

proc select( $\mathcal{S}, \mathcal{X}$ ) =
   $S_{BEST} := null; m_{BEST} := 0;$ 
  for  $S \in \mathcal{S}$  do
     $m := M(S);$ 
    if  $m > m_{BEST}$ 
       $S_{BEST} := S; m_{BEST} := m;$ 
      output( $S_{BEST}$ );
      remove  $S_{BEST}$  from  $\mathcal{S}$ ;
      split  $\mathcal{X}$  into subsets  $\mathcal{X}_1, \mathcal{X}_2, \dots$ , based on  $S_{BEST}$ ;
      if |subsets|  $\leq 1$ 
        return;
      fi
      for each subset  $\mathcal{X}_k$  do
        select( $\mathcal{S}, \mathcal{X}_k$ );
      od
    fi
  od.

```

4.2 Application to the Biological Case Study

The biological community has built over the years a comprehensive collection of resources to archive experimental data. As molecular biology became a more data intensive field with the advent of DNA microarrays, came the need to store not only measurements but also ancillary annotation describing experimental conditions and set up, thus ensuring a metadata core always shipped with the data set. The

Table 1. A fragment of the input document passed to the taxonomy generation algorithm. Schemes are grouped by common names preceding the semi-colon, e.g. *S1:On Material* refers to schema 1 and *On Material* is the classification.

Process Name	Occurrences	S1:Material	S1:Data	...	S6:in vitro	S6:in vivo	S6:in silico
labeling	390811	1	1
nucleic acid extr.	350267	1	1
hybridization	345671	1	1
feature extr.	267044	...	1	1
bioassay data trans.	176347	1	1
grow	116194	1	1	...
pool	68004	1	1
...

sizes of microarray databases (GEO, AE) constitute prime resources for evaluating and testing our approach [2, 1]. The content of ArrayExpress was therefore accessed obtaining data via parallel calls, converting 21000 experiments and associated experimental metadata to ISA-Tab[46, 50]. This step provided a harmonized format in which to represent not just transcriptomic data, but also genomic, proteomic, metabolomic and other classical experiment types.

Given the data sets, the code analyzes the content of these directories to determine the processes and IOs which exist within the experiments, and the number of times they occur. The analysis revealed 61 processes (a small number resulting from the homogeneity of the database where analysis techniques are targeted primarily towards DNA microarrays) with 1,845,089 occurrences and 8,223 properties of the sample of which 3492 were deemed to be IOs with 486,353 occurrences.

Several distinct, empirical but meaningful, categorizations were devised encompassing a number of facets defining the properties of the experimental process, either in terms of its participants or in terms of key process properties. Classifications, based on features such as the nature of process participants, the granularity scale, the nature of experiments and common types a treatments applied in biological experiments were developed. Some classifications were somewhat informed by the overall assumptions ISA model relies upon, where nodes can be either material or data files and where edges are processes acting upon those nodes. Others resulted from applying a small number of axioms discovered through consultations with domain experts. There were 6 classification schemes in all with a total of 23 sub-classifications, these are detailed in figure 3. Table 4.2 shows a snapshot of the input file used in the classification algorithm.

The schemes are not tied to any existing taxonomic tree, may not be orthogonal and/or may be redundant with respect to others. The development of the classifications was not the result of application of any knowledge engineering methods and there is no ontological commitment, although in theory the classification names used could be derived from an ontological framework. The reason for not relying on an ontology in the first place was risk mitigation, where use would almost certainly result in an unbalanced tree as occurrence counts for the terms would not be considered, resulting in creation of many glyphs that would never be used.

The creation of the tree shown in Fig. 3 was a largely iterative quest with continuous involvement of domain experts checking to ensure that the classifications assigned were meaningful. From early versions of the classification matrix creation, classifications were removed, added or merged in consultation with those who knew the data domain. For example, in S2 there were sub-classifications in *Genetic Modification & Labeling* which are types of *Material Combination*. Therefore, these sub-classifications were removed since it would be technically and semantically incorrect to keep them. These early interactions emphasized the importance of having domain experts in the loop, otherwise classifications and subsequent glyphs would not be as well created as they could be.

The algorithm managed to correctly place the classifications where they were expected (checks were made by hand to ensure that the algorithm performed well). Schema S6 (*in vitro*, *in silico* and *in vivo*) was the best top level classification due to its overall fitness metric of 3.08 with individual metrics found to be: M1 = 1.0; M2 = 1.0; M3 = 0.9; and M4 = 0.18. S6 was closely followed by S1 (on material and on

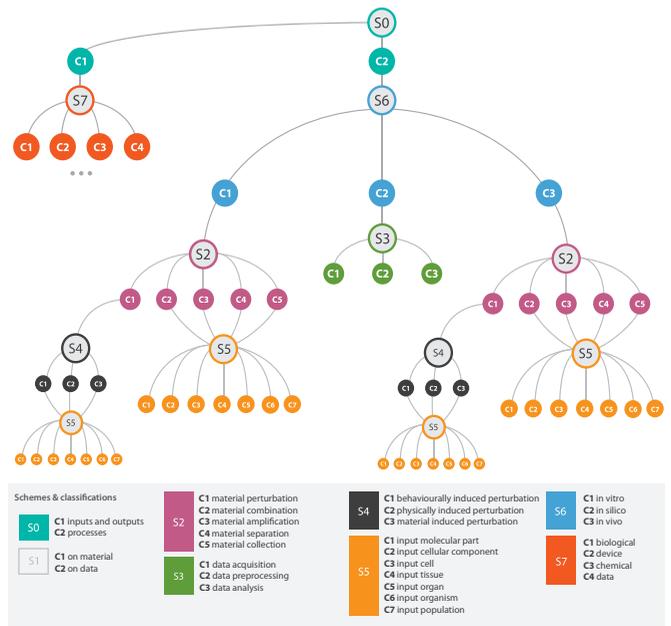


Fig. 3. The classification algorithm arranged the classification schemes in the above order.

data), however after the top level separation, S1 became redundant as a result of the top level implicitly making a split on data (in silico) and materials (in vitro and in vivo). The algorithm successfully detected this redundancy; hence S1 is absent from Fig. 3.

Although the algorithm performed as expected, a possible criticism of the algorithm could be highlighted in the placement of S5 below S4. S5 could have been placed above S4 in line with the observation that, for the majority of cases, S5 was selected over S4. The reason S5 wasn't used to sub-classify C1 of S4 is due to its number of classifications (7) being greater than that of S4 (3), therefore S4 was selected primarily on this metric even though the sub-tree balance of S5 (0.8) was slightly better than that of S4 (0.68). To modify this behavior, it was necessary to involve domain experts in refining the tree based on their domain-specific knowledge and case-specific requirements.

5 VISUAL ENCODING

5.1 Perceptual Guidance

A *glyph* is a small visual object composed of a number of *visual channels* which can be used independently as well as constructively to depict attributes of a data record. Glyphs are of a type of visual signs that can make use of visual features of other types of signs, such as icons, indices and symbols.

Although there are several books on sign design (e.g., [6, 4]), they focus on signage in public space, and offer empirical guidance on a large number of issues including standardization, size, location, illumination and so on. Many of these issues are not quite relevant to the need for visualizing the workflows of biological experiments. Here, we draw our design principles mainly from findings in perception, especially in the area of visual search [54, 43].

While the extensive use of signs, icons and pictograms in everyday life reflects their usefulness and effectiveness, several perceptual studies also directly or indirectly confirmed their perceptual and cognitive merits. For example, Franks and Bransford's study on transformation of prototypes [16] suggested that humans can learn to recognize glyphs by rules consciously as well as unconsciously. The presence of iconic memory [53] may facilitate rapid comparison between glyphs in the same display, whereas it is less so for texts.

Guideline on Semantic Relevance. Bertin [7] classified visual channels (which he referred to as retinal variables) into two categories, planar (location) and retinal (size, color, shape, orientation, texture and brightness). Bertin proposed four semantic criteria for determin-

Table 2. Perceptual strength of different visual channels based on levels of organization studied by Bertin[7] and Green[17], 'popout' effect studied extensively in psychology [61, 15, 31, 7, 17, 63, 55, 38, 39], and hierarchy effect studied by Navon [35] and Love [30]. The summary strength of each channel is estimated and it should be considered in conjunction with application-specific conventions and metaphors.

Regions	Visual Channels	Levels of Organization				Popout Effect	Hierarchy Effect	Summary Strength	Convention & Metaphor
		Associative	Selective	Ordered	Quantitative				
a) Main	1. Color	Yes	Yes	Yes		*****	Strong	*****	application specific
	2. Size		Yes	Yes	Yes	***		***	
	3. Shape	Yes	Yes			***		***	
	4. Orientation	Yes	Yes			***		***	
	5. Texture	Yes	Yes	Yes		**		**	
b) Supplementary	6-10 (same as 1-5)						Medium		
	11. Planar	Yes	Yes		Yes	***		***	
c) Interior	Contains (a+b))						Low		

ing the suitability of different channels in representing certain types of information. These semantic criteria are: *associative*, *selective*, *ordered* and *quantitative*. These criteria are important guidelines, though there has been disagreement in the literature as to how individual visual channels are judged. For example, Bertin considered shape as a non-selective variable. Research has shown that shapes such as filled rectangles, circles and triangles do not allow the human visual system to identify one shape from another effectively in a rapid action when they form some global structures [30, 35], (they have poor "pop-out" effect). However, the omission of all shapes as a selective visual channel has been challenged, for example, by [56, 57, 17], who show practice and familiarity can support selectivity with almost any shape.

Guideline on Channel Composition. As a glyph is likely to feature a number of visual channels, the constructive composition may affect how individual channels are perceived. A rich collection of literature on integral and separable dimensions shows that the combined dissimilarity of closely integrated visual channels exhibits Euclidean distance $\sqrt{d_a^2 + d_b^2}$ [26, 18], whereas that of separable visual channels exhibits city-block distance $d_a + d_b$ [10, 51]. The latter is more cost-effective than the former in rule-based encoding of multi-faceted concepts, therefore effective glyph design should encompass a non-conflicting set of separable retinal variables.

Guideline on Pop-out Effects. Many classic studies in perception also established the "power" of different visual channels in terms of *pop-out effect* (pre-attentive search), and fixation (during attentive search)[19]. The *pop-out effect* is one which allows identification of a target within a few nanoseconds of initial exposure to the visual search space. A result of several milestone studies focusing on observed response times, the ordering of the four commonly used visual channels follows the consensus: color \prec size \prec shape \prec orientation (e.g., [61, 44, 48]). The symbol \prec reads as *precedes*. However, the strength of color over the other three channels is generally much more noticeable.

Guideline on Visual Hierarchy. *Visual hierarchy*, with which the environment and objects around us are arranged is a well documented theoretical framework [38, 35, 30, 24, 5]. However, the literature contains a debate over the ways in which the visual system traverses this hierarchy. There are four possible ways: top-down (also called global processing) [35]; bottom-up (also called local processing); middle-out [24]; and salient features (e.g., *edges*, *points*, *colors*) [49]. Because glyphs are relatively small in comparison with a whole visualization, we consider at such a "localized level", the top-down and salient features may play more significant roles. The top-down assumption suggests that when consider a glyph in isolation, its global feature will affect visual search more than its local features. Salient features are partly addressed by the pop-out effects.

Based on the above-mentioned perceptual guidance, we considered a generic glyph design template for this work as shown in Fig. 4(a). The template divides a glyph into three regions, namely *main body*, *exterior* and *interior*. In practice, each region can be divided into 2 or more sub-regions (e.g., a twin body glyph or 4 exterior sections), it is convenient to consider the three regions in abstraction. The separation of these regions facilitates the basic separation of visual channels based on the composition guideline, while allowing us to consider them individually according to the hierarchy guideline.

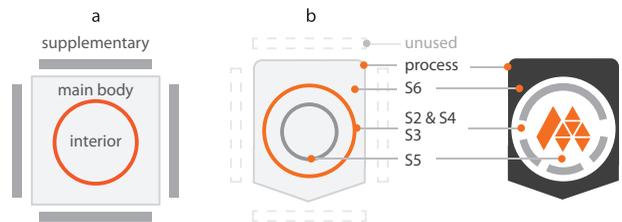


Fig. 4. (a) The glyph template with a main body, an interior region and an exterior region (consisting of 4 sections). (b) Our final design makes use of the main body and interior region. The exterior is reserved for future extension.

In theory, the exterior and interior regions may also be divided into sub-regions in a recursive fashion though in practice this is tightly constrained or discouraged by the very limited display resolution typically available for glyphs. Similar to the design convention for icons, pictograms are normally featured in the interior region. The exterior region may be further divided in four sections (top, bottom, left and right). If glyphs will be connected to form a network or graph (as in this work), the use of these four sections has to take into account the possible incoming and outgoing connections.

Table 2 summarizes relative merits of some of the most commonly-used visual channels in different regions according to Bertin's categorization, pop-out effects and hierarchy effects. We estimated the overall discriminating capacity of each channel by using a summary rating in the penultimate column. We also recognized the importance of the conventions and metaphors in an application. We will discuss visual metaphors further in Section 5.2. We added the last column to highlight the necessity to consider these in a design process.

It is difficult to establish an accurate ranking order of different visual channels by taking all perceptual effects into account in a quantitative manner. The state of the art in perception research is yet to provide all evidence needed for a full and conclusive analysis. Nevertheless, Table 2 can serve a qualitative guidance to glyph designs in this work, providing an ordering of visual channels in parallel with the ordering of categorization schemes discussed in Section 4.

5.2 Mapping Taxonomy to Visual Channels

For each scheme in the taxonomic tree as shown in Fig. 3, we propose a number of design options. Fig. 5 shows some examples of the proposed designs. For example, the first column shows the use of colors to encode the classes of a categorization scheme. The second column shows the use abstract shapes. Some options convey an abstraction from pictorial representations of classes, and in other cases, we try to establish a metaphoric association between a visual channel and a biological categorization.

Metaphoric visual representations enable domain-specific encoding using "natural mapping" [52, 36]. This natural mapping can make it easier for users to infer meaning from the glyph with less effort required to learn and remember them [33]. A recent study showed that visual metaphors can aid memorization of the information depicted in a visualization [8]. However, the same study also showed that visually realistic metaphors (those with a lot of detail) may have a negative

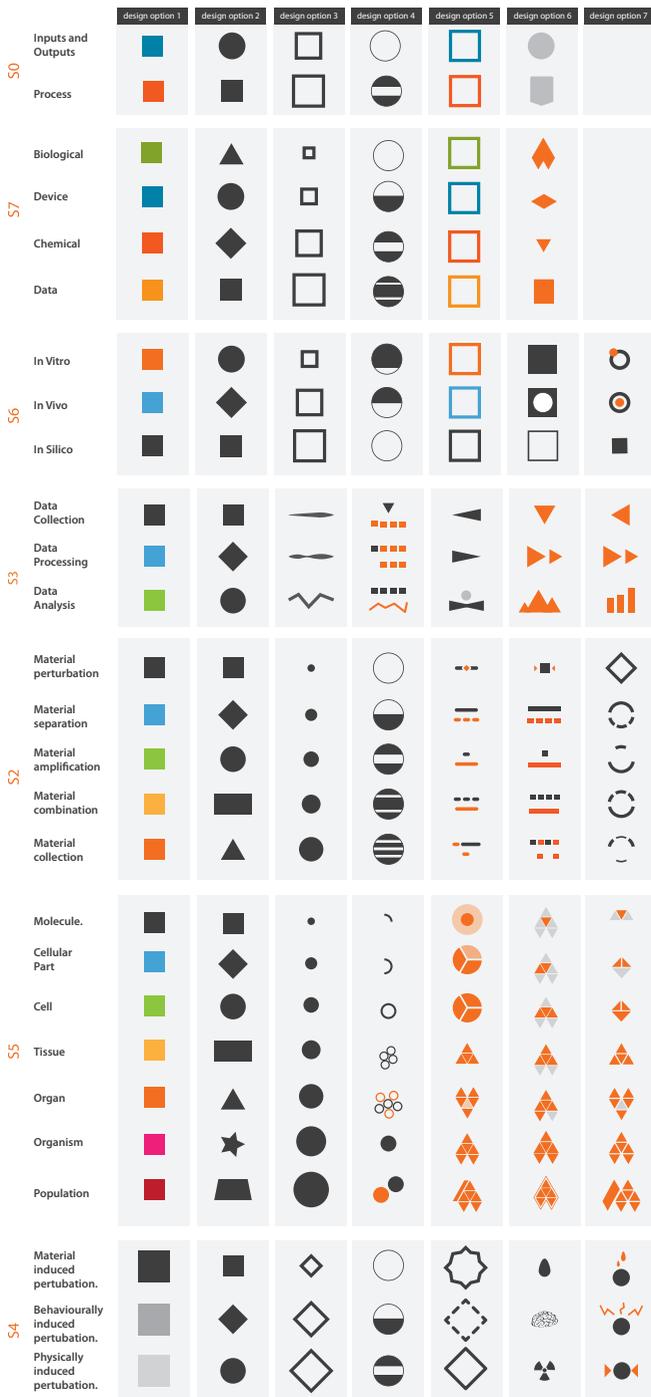


Fig. 5. Experimenting with visual channels: an overview of the various design options available for use in representing the different classifications. Most schemes for IO categorization are not shown due to space restriction.

impact on performance in visual search. Moreover, realistic visual metaphors require a higher pixel resolution, and would lose their discriminating capacity in low resolution conditions.

Based on the design options shown in Fig. 5, we followed the taxonomic tree in Fig. 3 and identified the best option for each scheme in a hierarchical manner. The evaluation criteria include:

- the discriminating capacities of different channels (Table 2);
- metaphoric capacity for aiding learning and remembering;
- potential conflicts, including spatial, perceptual and metaphoric conflicts, with visual channels that have already been assigned to

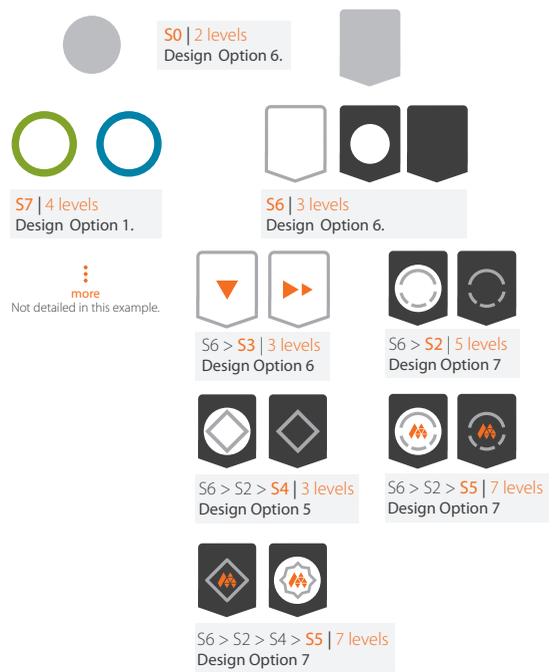


Fig. 6. Formation of the final glyph design. Top level items require the greatest visual power. It is important to be able to distinguish each of the processes based on their parent level in the taxonomy. Distinguishable global shapes provide the difference between IOs (circles) and processes (squares with pointed bottom to indicate directionality).

other schemes in the tree;

- encoding costs in terms of requirement for pixel resolution.

This process normally takes a few iterations, during which new design options and new metaphoric abstractions and associations are sometimes proposed.

Based on the hierarchy in Fig. 3, we considered to use color and shape options for S0 (IOs vs. processes) and S7 (four classes of IOs). As introducing 4 main body shapes to encode IOs is not an effective mapping for learning and remembering, we decided to assign outline color of the main body to encode S7, and use two basic shapes, circle and pentagon (a rectangle with a pointer to show workflow direction) to encode S0. Since the main body or the pentagon will only be colored in black or white, S0 also implicitly encoded in using color symbolism, that is, color for IOs and black/white for processes. In effect, S0 is encoded using two visual channels. This redundancy can serve as an error detection in visualization [12].

Fig. 6 shows each of the five schemes for the process subtree, and the design option chosen from Fig. 5. Below we discuss our reasoning for selecting each of the design options for each scheme in the taxonomy.

S6: Process environment. The taxonomic tree suggested a high priority for visual mapping, which is consistent with the domain experts’ intuition. We took advantage that black color was not used by S7 for IOs, we assigned a white background to *in silico/in computer* (related to computational processes), and black to *in vivo/in living* and *in vitro/in glass* (related to materials). Further more, we made use of a shape-based metaphor, fully-filled background for *in vivo* (whole organism), and black background with white cut out for *in vitro* (component of an organism). Together, this was given in Fig. 5 as design option 6. We maintained the overall appearance of the main body of process glyphs in black and white to avoid potential clash with material glyphs.

S2: Types of Material Manipulation. S2 has 5 classes, and we adopted design option 7, which employs visual metaphors that encapsulate strong domain-specific meanings. For example, visual symbol



Fig. 7. Visual representations of the 7 classes in S5.

for the *material amplification* class depicts a small segment becoming a large segment.

S4: Type of Experimental Perturbation. S4 defines 3 further subclasses of the *material perturbation* class of S2. Due to the low number of subclasses, we made use of line styles to modify the diamond shape of the *material perturbation*. We made metaphoric association of three line styles as: “solid” line for physically induced perturbation; dash line, a common metaphor for uncertainty and unpredictability, for behaviorally induced perturbation; and wavy line, which is closer to a circle (a visual signature for IOs), for material induced perturbation.

S5: Levels of Material Granularity. This scheme has 7 subclasses (molecule, cellular part, cell, tissue, organ, organism and population), and finding a suitable visual channel was not straightforward. A simple approach would be to use colors to fill the interior of the glyph, or to use some shape-based encoding. After consulting the domain experts, these two options were ruled out. In fact, domain experts preferred some pictograms to represent the 7 subclasses. After considering a number of more realistic drawings, we found that it was not easy to create realistic representations that can differentiate all subclasses (e.g., cellular part vs. cell; organ vs. organism). We designed a special set of icons as shown in Fig. 7 with three visual channels to aid learning, memorization and recognition. The first visual channel is the overall shape and orientation. The second visual channel is metaphoric abstraction and association. For example, the shape of cell part indicates a portion of a cell. Tissue is associated with a patch, organ with an abstract heart shape, organism with an abstract human, population with two abstract humans. The third visual channel is the number of orange sub-shapes, representing the levels 1-7.

S3: Types of Data Manipulation. S3 defines a 3 further subclasses of the *in silico* class of S6. We use three abstract pictograms to represent data capture, processing and analysis. The encoding makes use of the difference in overall shape, orientation, and number of triangles to aid learning, memorization and recognition. Note that these shapes will not be confused with those for S5 as the white background of the main body provides a distinct context of computing rather than IOs.

Following mapping of all schemes to design options, creation of all glyphs is a straightforward process. Fig. 8 shows all variations of the glyphs associated with the process sub-tree.

We introduced a “crush” test for the designed glyphs by scaling it to different pixel resolutions. Fig. 9 shows some example glyphs display at varying resolutions. One can comfortably see all details at the 40×40 level. At the 10×10 level, one can observe the visual signature of *in vivo* and *in vitro*. Even at the lowest level (5×5), one can still differentiate the two glyphs.

6 WORKFLOW VISUALIZATION AND ISA INTEGRATION

In order to visualize workflows of biological experiments, we had to address the following technical issues: 1) mapping from a name in the metadata to a glyph; 2) creating a workflow visualization with both node placement and connection display; 3) developing a prototype tool for a practical environment such as the ISA tools framework.

The mapping from concept name is achieved by a look-up table which is built from the matrix used in formulating the taxonomic tree and implemented as a tab delimited file. Each concept name is mapped to a text tag encoding the traversal path from the root of the tree to the leaf node corresponding to the name. The tag includes identifiers of the schemes and classes encountered. With the path, a glyph can be constructed dynamically from the pre-defined visual mapping as described in the previous section. The look-up table also enables storage of pre-rendered glyphs in an image format.

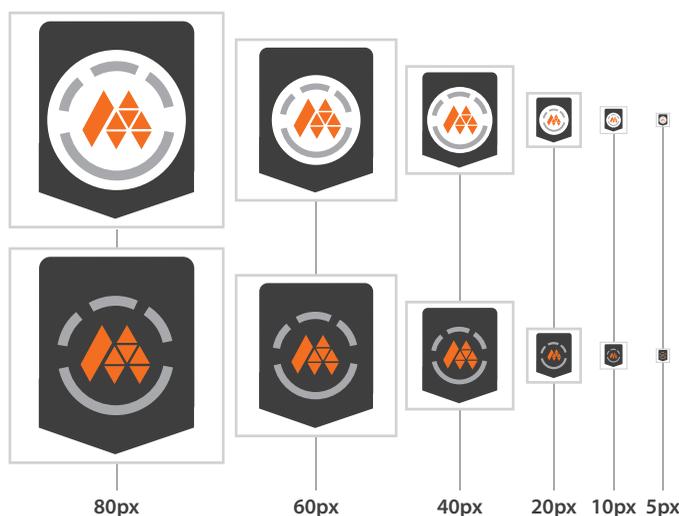


Fig. 9. A “crush” test of glyphs to evaluate the discriminating capacity of various visual channels at different resolutions.

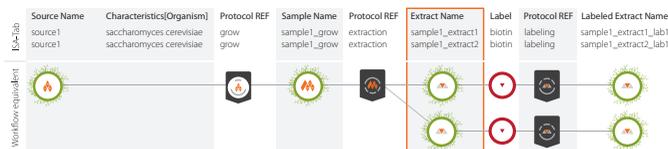


Fig. 10. A workflow is recorded in text form within the ISA-Tab format. Our software translates it to glyph-based visualization. A branching event is automatically detected during the translation.

To generate the workflow, we made use of the layout algorithm available within the Prefuse visualization framework [20]. This framework also brings with it functionality such as panning, zooming and filtering which bring more interactivity to the user and making navigation through large collections of workflows easier. The only requirement for use of Prefuse was a minimal amount of Java code to create the user interface coupled with creation of an XML file format native to Prefuse for representation of the tree structure. This XML format is configurable, we have configured it to contain: *node type* (e.g., process), *node name* (e.g., labeling) and an image to be rendered, which is assigned using a look-up operation through the above mentioned tab delimited mapping file. The order of the XML elements within this file has direct implications on the order these elements are displayed in. Within the ISA-Tab format, there is an implicit time element found in the ordering of the columns in the study sample and assay files. This can be used to construct the XML elements through near direct mappings of processes and IOs, a workflow which is illustrated in Fig. 10. Recognizing branching events is an important part of workflow visualization as illustrated in 10. In software, these branch events can be identified when the preceding nodes of a process have the same names while the succeeding output nodes are different. Fig. 10 highlights one such case, where branching occurs after extraction of 3 materials from one sample.

Our software reads text-based ISA-Tab files as illustrated to create the XML notations required by Prefuse for rendering the experiment workflow illustrated in Fig. 1(b).

The software is implemented as a Java application capable of processing ISA-Tab files to create experimental workflows using glyphs. For the purposes of broad dissemination in the near future, we have integrated the workflow visualization directly inside ISAcreeator, a Java desktop application for creating and editing ISA-Tab files. Furthermore, the standalone workflow visualization shown in figures 11a and b makes use of the same ISA-Tab parser as available within ISAcreeator.

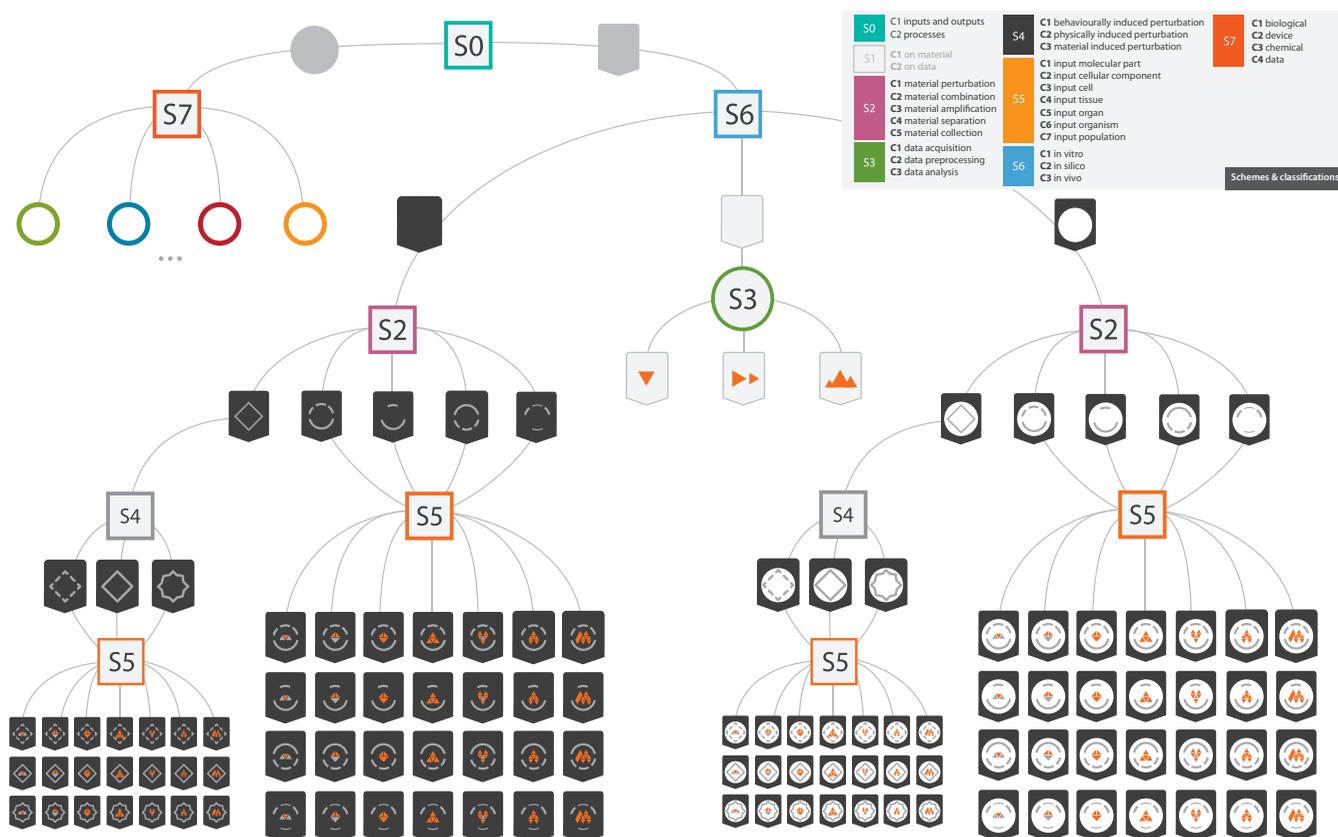


Fig. 8. Overview of the taxonomy based glyph designs in the context of biodomain.

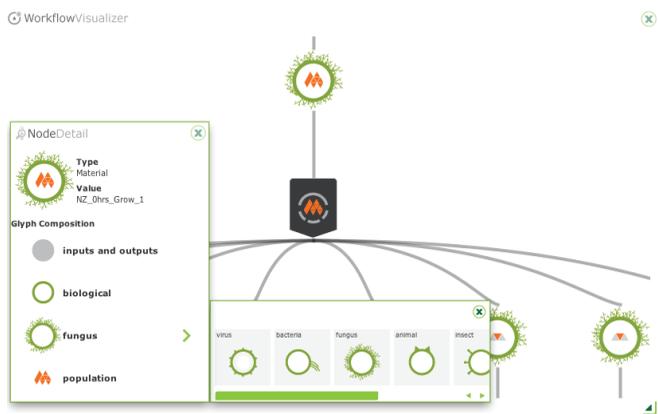


Fig. 11. User can interact with a workflow to view the text descriptions of individual glyphs in pop-up windows, which are also dynamic legends showing how the concept is categorized and how the corresponding glyph is decomposed into different elementary visual representations. Users can also use such a legend to find out other classes in a categorization scheme.

ator. The integration involved the development of user interface elements for ISAcreator software to access the workflow visualization. From the spreadsheet view, users may highlight rows (assays), right click and select an option named “View workflow for assays” to visualize the entire processing pipeline for the corresponding samples as shown in Fig. 1(b).

7 CONCLUSIONS AND FUTURE WORK

Although it was established in psychology that people can learn rules of glyph encoding consciously as well as unconsciously [16], we do

not expect users to learn and remember these glyphs without help. We thus allow users to find detailed text descriptions in pop-up windows interactively. As shown in Fig. 11, not only do these windows provide names of processes and IOs, they also serve as dynamic legends, showing how each glyph is decomposed into different elements corresponding to categorization schemes. One can also select a component to view all classes in a scheme.

In this manuscript, we presented a systematic approach for glyph design by using taxonomy as a guide. We demonstrated our approach by analyzing the content of a biological database and showed how our tree-building algorithm could be used to take a large collection of concepts and generate a taxonomic tree, which provides an ordering of a set of categorization schemes (i.e., attribute dimensions). This enabled us to draw a parallel between the ordering of attribute dimensions with the ordering of visual channels compiled from the perception and visualization literature. We involved domain experts in refining the tree as well as in creating metaphoric abstraction and association for glyphs. This work was followed by development of a prototype tool for glyph-based visualization of experimental design workflows as found in biology. The prototype is integrated in the ISA suite of tools to be disseminated to users.

We plan to further the present work to include the provision of “macros” in workflows in order to make graphs more compact and create dynamic web-based rendering of the workflows using vector graphics. We also intend to carry out field-based evaluation through ISAcommons, the user community of ISA-tools framework. Like all potential diagrammatic schemes, we expect that it will take many iterations before it can become a standard. Nevertheless, the development of an online visualization tool will make such a process more efficient.

ACKNOWLEDGMENTS

This work was supported by funding from the BBSRC and NERC, grants BB/I000917/1 and BB/I025840/1.

REFERENCES

- [1] Arrayexpress. <http://www.ebi.ac.uk/arrayexpress/>.
- [2] Gene expression omnibus (geo). <http://www.ncbi.nlm.nih.gov/geo/>.
- [3] Graphviz. <http://graphviz.org/>.
- [4] R. Abdullah and R. Hümbler. *Pictograms, Icons & Signs: A guide to Information Graphics*. Thames & Hudson, 2006.
- [5] M. Bar. Visual objects in context. *Nature reviews. Neuroscience*, 5(8):617–629, 2004.
- [6] P. Barker and J. Fraser. Sign design guide: a guide to inclusive signage. *JMU and Sign Design Society*, 2000.
- [7] J. Bertin. *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press, 1983.
- [8] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An empirical study on using visual embellishments in visualization. *to appear in IEEE Transactions on Visualization and Computer Graphics*, 2012.
- [9] R. Bourqui and M. Westenberg. Visualizing temporal dynamics at the genomic and metabolic level. *Proc. 13th Int. Conf. Information Visualization*, pages 317–322, 2009.
- [10] B. Burns, B. E. Shepp, D. McDonough, and W. K. Wiener-Ehrlich. The relation between stimulus analyzability and perceived dimensional structure. *The Psychology of Learning and Motivation*, pages 77–115, 1978.
- [11] Y. Cai, M. L. Wilson, and J. Peccoud. Genocad for igem: a grammatical approach to the design of standard-compliant constructs. *Nucleic Acids Res*, 38(8):2637–44, May 2010.
- [12] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, 2010.
- [13] M. Chuah and S. Eick. Information rich glyphs for software management data. *IEEE Computer Graphics and Applications*, 18:24–29, 1998.
- [14] T. Czauderna, C. Klukas, and F. Schreiber. Editing, validating and translating of sbgn maps. *Bioinformatics*, 26(18):2340–2341, 2010.
- [15] J. Duncan and G. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433–491, 1989.
- [16] J. Franks and J. Bransford. Abstraction of visual patterns. *Journal of experimental psychology*, 90(1):65–139, 1971.
- [17] M. Green. Toward a perceptual science of multidimensional data visualization: Bertin and beyond. *ERGO/GERO Human Factors Science*, 1998.
- [18] S. Handelt and S. Imai. The free classification of analyzable and unanalyzable stimuli. *Attention, Perception, & Psychophysics*, 12(1):108–224, 1972.
- [19] C. G. Healey and J. T. Enns. Attention and visual memory attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2011.
- [20] J. Heer, S. K. Card, and J. A. Landay. prefuse: A toolkit for interactive information visualization. In *Proc. ACM CHI*, pages 421–430, 2005.
- [21] K. Hemenway. Psychological issues in the use of icons in command menus. In *Proc. ACM CHI*, pages 20–23, 1982.
- [22] B. H. Junker, C. Klukas, and F. Schreiber. Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 2006.
- [23] A. Karve and M. Gleicher. Glyph-based overviews of large datasets in structural bioinformatics. In *Proc. 11th Int. Conf. Information Visualization, Supplements*, pages 1–6, 2007.
- [24] R. Kinchla and J. Wolfe. The order of visual processing: "top-down," "bottom-up", or "middle-out". *Perception & psychophysics*, 25(3), 1979.
- [25] G. Kindlmann and C.-F. Westin. Diffusion tensor visualization with glyph packing. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1329–1335, 2006.
- [26] D. Krantz and A. Tversky. Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, 12(1), 1975.
- [27] R. Krishnapuram and K. Kumnamuru. Automatic taxonomy generation: Issues and possibilities. In *Proc. IFSA 2003, LNCS*, pages 184–195, 2003.
- [28] N. Le Novère, M. Hucka, H. Mi, S. Moodie, and *et al.* The systems biology graphical notation. *Nat Biotechnol*, 27(8), 2009.
- [29] J. P. Lewis and R. Rosenholtz. VisualIDs: Automatic distinctive icons for desktop interfaces. *ACM Transactions on Graphics*, 23(3):416–423, 2004.
- [30] B. Love, J. Rouder, and E. Wisniewski. A structural account of global and local processing. *Cognitive psychology*, 38(2):291–607, 1999.
- [31] S. Luck and S. Hillyard. Electrophysiological correlates of feature analysis during visual search. *Psychophysiology*, 31(3):291–308, 1994.
- [32] E. Maguire. Taxonomy in biology and visualization. <http://isatab.sourceforge.net/docs/publications/Taxonomy.pdf>, Jan 2012.
- [33] S. McDougall, O. De Bruijn, and M. Curry. Exploring the effects of icon characteristics on user performance: The role of icon concreteness, complexity, and distinctiveness. *Journal of Experimental Psychology: Applied*, 6(4), 2000.
- [34] P. Muter and D. Mayson. The role of graphics in item selection from menus. *Behaviour and Information Technology*, 5:89–95, 1986.
- [35] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–736, 1977.
- [36] D. A. Norman. *The Design of Everyday Things*. Basic Books, 2002.
- [37] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *NAR*, 27(1), 1999.
- [38] S. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4):441–915, 1977.
- [39] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1), 2002.
- [40] J. K. Patel and C. B. Read. *Handbook of the Normal Distribution*. Marcel Dekker, 2nd edition, 1996.
- [41] J. W. Pellegrino, R. R. Rosinski, H. L. Chiesi, and A. Siegel. Picture-word differences in decision latency: An analysis of single and dual memory models. *Memory and Cognition*, 5:383–396, 1977.
- [42] F. Post, T. Walsum, F. Post, and D. Silver. Iconic techniques for feature visualization. *Proc. IEEE Visualization*, pages 288–295, 1995.
- [43] P. Quinlan. Visual feature integration theory: past, present, and future. *Psychological bulletin*, 129(5):643–716, 2003.
- [44] P. Quinlan and G. Humphreys. Visual search for targets defined by combinations of color, shape, and size: an examination of the task constraints on feature and conjunction searches. *Perception & psychophysics*, 41(5):455–527, 1987.
- [45] W. Ribarsky, E. Ayers, and J. Eble. Glyphmaker: creating customized visualizations of complex data. *Computer*, 27:57–64, 1994.
- [46] P. Rocca-Serra, E. Maguire, S.-A. Sansone, and *et al.* Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18), 2010.
- [47] Rohrer and *et al.* The shape of shakespeare: visualizing text using implicit surfaces. In *Proc. IEEE Visualization*, pages 121–129, 1998.
- [48] T. Ropinski, S. Oeltze, and B. Preim. Survey of glyph-based visualization techniques for spatial multivariate medical data. *Computers & Graphics*, 35(2):392 – 401, 2011.
- [49] D. E. Rumelhart. A multicomponent theory of the perception of briefly exposed visual displays. *Journal of Math. Psych.*, 7(2):191–218, 1970.
- [50] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, and *et al.* Toward interoperable bioscience data. *Nat Genet*, 44(2):121–6, 2012.
- [51] R. Shepard. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1(1):54–141, 1964.
- [52] H. Siirtola. The effect of data-relatedness in interactive glyphs. In *Proc. 9th Int. Conf. Information Visualization*, 2002.
- [53] G. Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29, 1960.
- [54] K. T. Spöhr and S. W. Lehmkuhle. *Visual Information Processing*. W. H. Freeman & Company, 1982.
- [55] A. Treisman. Focused attention in the perception and retrieval of multidimensional stimuli. *Attention, Perception, & Psychophysics*, 22(1), 1977.
- [56] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychol Rev*, 95(1):15–48, Jan 1988.
- [57] Q. Wang, P. Cavanagh, and M. Green. Familiarity and pop-out in visual search. *Percept Psychophys*, 56(5):495–500, Nov 1994.
- [58] M. O. Ward. Multivariate data glyphs: Principles and practice. In *Handbook of Data Visualization*, pages 179–198, 2008.
- [59] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2004.
- [60] S. Wiedenbeck. The use of icons and labels in an end user application program: an empirical study of learning and retention. *Behavior and Information Technology*, 18(2):68–82, 1999.
- [61] L. Williams. The effects of target specification on objects fixated during visual search. *Acta psychologica*, 27:355–415, 1967.
- [62] C. Wittenbrink, A. Pang, and S. Lodha. Glyphs for visualizing uncertainty in vector field. *IEEE Transactions on Visualization and Computer Graphics*, 2:266–279, 1996.
- [63] J. Wolfe, K. Cave, and S. Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of experimental psychology. Human perception and performance*, 15(3):419–452, 1989.