




Implementation Issues for a Highly Structured Research Report

Robert B. Allen^(✉) 

New York, NY, USA
rba@boballen.info

Abstract. We have proposed that scientific research reports should be constructed entirely of structured knowledge rather than text. In an earlier paper, we emphasized Research Designs as a framework for structured research reports and described how a structured implementation might be applied to Pasteur’s classic swan-neck flask experiment. In this paper, we examine some of the issues encountered in developing that implementation using dynamic models. For instance, we consider issues associated with modeling state transitions.

Keywords: Declarative programming · Direct representation · Hypothesis Model · Pasteur · Research Designs

1 Semantic Modeling for Research Reports

Over time scientific research reports have become increasingly structured. That trend has accelerated with automated and flexible data management and description. Among these are [8, 12] (see [11] for an overview). However, we have found no attempts to structure all aspects of research reports as our approach proposes [1–7].

Structured scientific research reports are arguments for the research claims. They are complex digital objects and should be implemented with comprehensive, standard vocabularies, based on general world knowledge, and previous research results. Potentially, the structured reports will become the foundation for a knowledge-based scientific digital library. The heart of a research report is interwoven sequences of transitions. In [6], transitions are described as state changes with rules or triggers that activate when given conditions are met. One sequence of transitions is the Research Procedure which is based on the Research Design [7]. For experimental Research Procedures, the triggers (manipulations) are the actions of the researcher. The other sequence comprises causal Hypothesis Models for the phenomenon under investigation. Potentially, Research Procedures will cause a sequence of transitions in the research environment (microworld) as predicted by one or another of the hypotheses. In [7] we described these two types of sequences as “yoked” in experiments.

Goal and Roadmap: There are many advantages to highly structured research reports. They would facilitate research claims validation, inferences, and interactive tutorial

explanations. This paper extends our initial proposal for a structured report of Pasteur’s swan-neck flask experiment [7] by examining implementation issues. We focus on developing robust representations and dynamic models. Our approach is related to object-oriented modeling; we model the interaction of objects as a dynamic simulation across time, with the research set in a microworld. In our implementation, the classes are based on the SUMO ontology [17] extended to cover the vocabulary required for the specific research domain. We do not focus on text mining. Nor do we focus on inference across large knowledgebases, and we do not at this point emphasize machine learning or the cognition of discovery [19].

Section 2 explores general issues for developing structured research reports. Section 3 focuses on developing a structured report for a somewhat simplified version of Pasteur’s Model swan neck flask experiment. This includes a software implementation of a Hypothesis Model. Section 4 discusses the potential for a broad knowledgebase of research reports and related materials.

2 Highly Structured Research Reports

A typical research report is a structured argumentation document. It has a Goal, Question, Strategy, Hypothesis, Procedures, Results, and Claim. The strength of the argument is measured by accepted standards of evidence given different research designs and methods. Together they provide a “warrant” for further acceptance of the Claim.

The Research Goal is the motivation for the research. It is often based on human needs and values such as extending the human lifespan or improving the human food supply. In most cases, the Research Question addresses a specific issue related to the Research Goal. Because determining causation is one of the central goals of science [9], the Research Question may be framed as finding rules for considers causal transitions. A variety of Research Designs has been explored for structuring the relationship between the hypotheses and the procedures [18]. Different designs (e.g., experimental, quasi-experimental, observational) support different strengths of inferences about the results based on the internal and external validity. Construct validity, which considers whether the constructs being examined match reality, is closely related to ontology. A paradigm shift is the result of the realignment of several related constructs [21].

The Hypothesis is a simple causal transition (e.g., “biogenesis”) while the Hypothesis Models a causal sequence. Typically, Hypothesis Models are not standard mechanisms [6] but ad hoc sequences applied to the specific research environment. The Procedure is the actions taken by the researcher. Because the Procedure and Hypothesis Models in experiments are interdependent, we say they are yoked [7]. Since experiments often require a highly controlled environment, the Procedure may include presets to the state of objects in the microworld.

Figure 1 and Fig. 2 illustrate the Procedure (red) and the Hypothesis Model (blue) in two contexts, experimental and observational. In both cases, the causal path of the Hypotheses Model moves from the independent variable (*IVhyp*) to the dependent variable (*DVhyp*). For experiments (Fig. 1), the manipulation by the researcher (*IVmanip*) is intended to trigger the *IVhyp* that might not be directly observed; there is

a causal sequence between IV_{manip} and IV_{hyp} (also blue). Similarly, DV_{hyp} may not be observed directly but only as reflected in DV_{obs} .

In observational studies (Fig. 2) there is a Procedure for the observations, but no specific manipulation. Rather, the focus is on the selection of conditions and variables and the interaction with other models. In some cases, observational research may examine specific IV_{obs} and DV_{obs} as shown in Fig. 2. In other cases, there may not be à priori expectations of the causal pathways. Here, we concentrate on experimental research; we plan to address observational research in later papers.

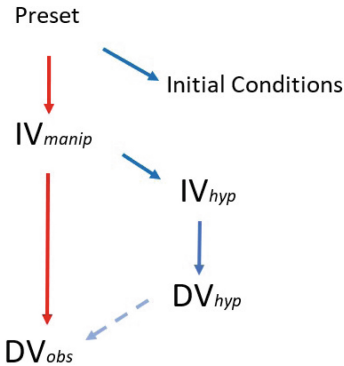


Fig. 1. Experimental research procedures (red) include a manipulation (IV_{manip}). This affects the Hypothesis Model (vertical blue). (Color figure online)

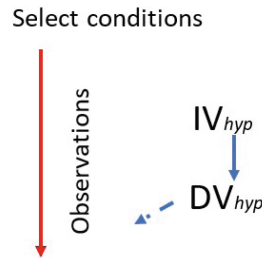


Fig. 2. Observational research may also explore Hypothesis Models but there is no manipulation thus, weakly yoked. (Color figure online)

In addition to the framework described above for the Procedures and Hypothesis Models, the obtained results need to be modeled. Planned comparisons apply specific tests to the observations to evaluate the hypotheses [4]. Other comparisons may be developed as appropriate. In most cases, the Claim reflects the initial research question and hypothesis although extended Questions and the Claims can be asserted at different levels of generality (see “Structured Annotations” below).

The myExperiment project [10] is a collection of community-deposited workflows based on Taverna. Our approach extends that project’s approach by pairing structured workflows with structured hypothesis (outcome) models along with adding structured Goals, Questions, Results, and Claims.

3 Pasteur’s Swan-Neck Flask Experiment

Background and Framework: Aristotle proposed that life can arise from nonliving matter; he believed that air contains vital heat, which causes the development of new organisms. This process is known as Spontaneous Generation. By comparison, Germ Theory was based on biogenesis, that is that “life is required for the creation of life”. There was an extended debate and several studies between proponents of Spontaneous Generation and Germ Theory. Around 1859, Pasteur [16] effectively settled the debate

with the elegant swan-neck flask experiment. The curved swan neck traps microbes and heavier particles but allows air to pass into the flask holding nutrient broth. Later, the flask is tilted so that the broth reaches the curve with the trapped microbes. Pasteur found that microbes grew in the broth only after the flask had been tilted. In [7], we outlined a framework for describing the swan-neck flask experiment; here, we examine issues in implementing that framework.

Ultimately the Research Goal is related to improving human health and food production by understanding more about microbes. While desirable microbes can produce wine and dairy products, undesirable microbes can create spoilage. Similarly, microbes can be beneficial or adverse to human health. By understanding the lifecycle of microbes, potentially spoilage could be minimized and health improved. Based on the Goal, the Research Question could be framed as whether Spontaneous Generation or Biogenesis accounts for the origin of microbes.

Representations: Dynamic models require transitions that may involve changes in an attribute, in a process state, or in the relationship to other objects. For detailed models, the transition must be defined in combination with a specific object. In our approach, an object is defined as a Python class for the transition and the transition as a function within that class that is applied to the object.

The greatest challenges are in the representation of complex objects that change through time. For example, for multi-granular objects and their transitions. It is not feasible to model each molecule of air or broth or each microbe and its behavior. Rather, we model them as groups and maintain a general characterization of their properties. Representing small changes in the count of numerous objects (such as due to the reproduction of microbes) remains a challenge for qualitative modeling. Thus, we used pseudo counts for attributes such as the number of microbes (also for intervals of time).

The more granular level is not implemented although it may be acknowledged as part of an explanation. This is implemented as a declarative model but not strictly object modeling. Perhaps, the details could be implemented with multi-granular models (e.g., [15]). Or perhaps limited object models could be introduced as examples or scenarios for the more general process. For example, microbes are carried by air currents. When air is still, as it is in the neck of the flask, gravity dominates and microbes fall to the bottom of the neck. We model this competition of forces with qualitative force dynamics, a concept adapted from linguistics [20].

A second major set of challenges relates to representing locations in a microworld. We implemented subRegions within which compound entities (e.g., collections of microbes) could have their own state. The regions can be dynamic as the objects change. Moreover, there are nuanced interactions between the regions and the objects. For instance, as the flask is tilted and a portion of broth flows into the neck, do we need to reduce the broth spatial subRegion in the body of the flask? Because it was not significant for our purposes, we did not implement a change in the flaskBody microworld. A notation could be developed to address these modeling issues.

Hypothesis Models, Procedure, Results, Comparisons, and Claims: Pasteur's experiment had two conditions; one in which the flasks are tilted to allow microbes

to enter broth and one in which they are not tilted. Because there are two Hypotheses (biogenesis and spontaneous generation), there are four (2×2) Hypothesis Models. Here, we focus on only one of the four in detail because of space limitations but the others are analogous. The short version of the Hypothesis Model is that the introduction of microbes to the sterile broth causes fermentation of that broth. That can be extended by more specifics about the transitions of the objects as the experiment progresses. In other words, the Hypothesis Model is a causal chain.

In addition to the initial models, the full report needs to include the actual results as well as comparisons and claims based on them. For the swan-neck flask experiment, there could be comparisons across conditions [3, 7] and across time (before tilting versus after). The results support biogenesis and not the spontaneous generation (vital heat) hypothesis. In most cases, the Claim is based on one of the hypotheses. When unexpected results are obtained, a different claim could be developed. If we believe that the microbes in Pasteur's research environment are typical of other types of microbes, based on induction [13], we could assert the broad Claim that:

The reproduction of microbes is a necessary condition for the development of new microbes.

Our willingness to accept generalizations depends on factors such as the strength of the research results, whether we know of counter-examples, and whether there is a plausible mechanism to account for the results. These are issues of external validity. For Pasteur's experiment, we could ask how typical the microbes in Pasteur's flasks are of the broader population of microbes, which is an external validity challenge due to the "Interaction of the Causal Relationship of Units" (Table 3.2 [18]). Disputes about the claims can be represented with structured Toulmin-style argumentation.

Structured Annotations: The components described above form a framework for the research report. That framework can be extended for annotations that could include metadata and reasons for choices and explanations. Because we are developing dynamic models there are transitions with many nuances. While under normal circumstances, a given transition may be triggered across a range of conditions, for a given research scenario, those conditions may be greatly restricted. As the details of the models are refined, those conditions can be sharpened. A structured annotation for a transition typically would include: (a) associated objects and case roles [7], (b) necessary conditions (e.g., triggers, input rules), inputs, outputs, and side-effects, (c) purpose, and (d) potential sub-processes and more finely-grained representation.

Claims could be associated with different levels of confidence and include structured annotations for or against internal and external validity. For instance, the researcher might use a "check on the manipulation" to strengthen the case for internal validity. Internal validity errors occur when the researcher's action does not have the intended effect, see Table 2.4 in [18]. This check would compare the intended state of the microworld with its actual observed state following the manipulation. Tables 2.4, 3.1, and 3.2 in [18] can be used as an initial categorization for criticisms of the research.

Software Implementation: We developed a Python implementation for the Hypothesis Model described above. Specifically, for the condition where tilting the flask results in microbe growth. The first step is establishing the microworld and objects such as the flask located in it. These objects are initialized to the states needed for the research. For instance, the flasks are filled with broth, that broth is boiled, and the swan neck on the flask is created. Subregions include parts of the flask such as the flask-body broth, the air in the upper portion of the flask body, and parts of the flask neck. Potentially, spatial partitioning with scene graphs used in computer graphics worlds such as in computer games could be applied as a data structure. Next, the ongoing processes associated with the microworld are initiated. There are air currents that carry the microbes from the external air into the flask neck. However, inside the neck, the air is mostly still; gravity is the dominant force on the microbes so they settle to the floor of the neck [20].

Many of the objects in the simulation are complex; they change state and interact with other objects at different points. To keep track of the current state each object class has its own copy of a Python list of relationships that are replicated down the inheritance tree and updated when there is a state change. As shown below, when a generic flask class is specialized as a glass flask, “we add a tuple that specifies the “madeOf” attribute to the Python list of relationships”.

```
RelList.append(["attribute"],["always"],["madeOf"],["cGlass"],["comment"]])
```

Hypothesis Models are causal sequences, of events that are triggered by changes in the states of objects in the environment. This suggests programming asynchronous events with declarative programming (cf., [14]). While a declarative program might have been implemented with a blackboard and scheduler or with threading, we used a clock-tick-based rotation through the subregions of the microworld where the current state of the objects is evaluated and updated if the trigger conditions are satisfied. The program runs to completion and generates the expected response for the biogenesis hypothesis under the flask-tilt condition.

4 Discussion and Conclusion

Knowledgebase: While we have focused on individual research reports, we envision digital-library-like collections of reports that could be annotated, indexed, and cross-linked. As suggested in [3], traditional document citations could be replaced by linking claims along with structured justifications for the relevance of the claim links (categories of citations). Moreover, there could be structured review-style documents discussing the reliability (replicability) of the effects, integrating and comparing the claims from multiple studies, and discussing the development of theories based on the studies.

The repository would be associated with a rich ontology and other types of world knowledge. For instance, the rule that “boiling kills microbes” is common knowledge, although if needed it could be derived from principles such as the biochemistry of microbes and the effect of heat and cached in the knowledgebase. In addition, the knowledgebase would also include structured research methods as well specialization and applications.

Summary: Highly structured research reports would have several advantages over traditional text reports. They could provide rich linking and be the basis of text generation and tutoring at varying levels of detail. In addition, interactive interfaces could be developed for exploring the research reports. These interfaces could allow users to get overviews and drill down into details as desired. While models of scientific phenomena are often idealized [22], our models should extend previous approaches by allowing exploration across different levels of granularity.

There are many open questions about how best to structure research reports. We have focused on relatively simple qualitative models but richer quantitative modeling techniques could eventually be incorporated [15]. Model operators will need to be developed for these extensions to the techniques described above.

References

1. Allen, R.B.: Highly structured scientific publications. In: JCDL, pp. 472 (2007). <https://doi.org/10.1145/1255175.1255271>
2. Allen, R.B.: Model-oriented scientific research reports. D-Lib. Mag. (2011). <https://doi.org/10.1045/may2011-allen>
3. Allen, R.B.: Supporting structured browsing of research reports (2012). arXiv: 1209.0036
4. Allen, R.B.: Rich linking in a digital library of full-text scientific research reports. In: Columbia University Research Data Symposium (2013), PDF. <https://doi.org/10.7916/D8JM2JZ4>
5. Allen, R.B.: Rich semantic models and knowledgebases for highly-structured scientific communication (2017). arXiv: 1708.08423
6. Allen, R.B.: Issues for using semantic modeling to represent mechanisms, 2018, arXiv: 1812.11431
7. Allen, R.B.: Yoked flows for direct representation of scientific research. In: Digital Infrastructures for Scholarly Content Objects (DISCO) (2021), CEUR, pp. 2976–2983
8. Bechhofer, S., DeRoure, D., Gamble, M., Goble, C., Buchan, I.: Research objects: towards exchange and reuse of digital knowledge. Nat. Proc. **1** (2010). <https://doi.org/10.1038/npre.2010.4626.1>
9. Ben-Menahem, Y: Causation in Science. Princeton University Press (2018)
10. de Roure, D., Goble, C., Stevens, R.: The design and realisation of the MyExperiment virtual research environment for social sharing of workflows. Future Generation Comput. Syst. **25**, 561–567 (2009). <https://doi.org/10.1016/j.future.2008.06.010>
11. de Waard, A., Kircz, J.: Modeling scientific research articles – Shifting perspectives and persistent issues. In: ELPUB, Conference on Electronic Publishing, Toronto (2008). CiteSeer: 10.1.1.578.6751
12. de Waard, A., Tel, G.: The ABCDE format enabling semantic conference proceedings, SemWiki, Workshop on Semantic Wikis, Budva, Montenegro. CEUR: 206/paper8.pdf (2006)
13. Holland, J., Holyoak, K., Nisbett, R.E., Thagard, P.: Induction: Processes of Inference, Learning, and Discovery. MIT Press (1986)
14. Kuipers, B.J.: Qualitative simulation. Artificial Intell. **29**, 289–338 (1986)
15. Park, M., Fishwick, P.A., Lee, J.: Multimodeling. In: Fishwick, P.A. (ed.) Handbook of Dynamic System Modeling, Chapman, pp. 14.1–14.28 (2007)
16. Pasteur, L.: Sur les corpuscules organisés qui existent dans l’atmosphère: Examen de la doctrine des générations spontanées: Leçon Professée À la Société Chimique de Paris le 19 Mai 1861.

17. Pease, A.: *Ontology: A Practical Guide*. Articulate (2011)
18. Shadish, W.R., Cook, T.D., Campbell, D.T.: *Experimental and Quasi-experiment Designs for Generalized Causal Inference*. Houghton, Boston (2002)
19. Shrager, J., Langley, P. (eds.): *Computational Models of Scientific Discovery and Theory Formation*. Morgan-Kaufmann (1990)
20. Talmy, L.: Force dynamics in language and cognition. In: Talmy, L. (ed.), *Toward a Cognitive Semantics*. MIT Press, Cambridge (Cambridge)
21. Thagard, P.: *Conceptual Revolutions*. Princeton University Press (1992)
22. Weisberg, M.: Three kinds of idealization. *J. Philos.* 639–659 (2007). <https://doi.org/10.5840/jphil20071041240>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

