



Holistic models as an integrative infrastructure for scientific communication

Michael Knoblauch^{*}, Winfried Peters

School of Biological Sciences, Washington State University, Pullman, WA, 99164, USA

1. Background

The invention of the movable type printing press by Johannes Gutenberg (~1400–1468) fundamentally changed the ways in which knowledge could be preserved and accessed. The availability of printed books allowed information to be distributed widely, and had a profound influence on the intellectual developments that occurred in Europe in the Renaissance and Enlightenment epochs. While it was not unusual before the 20th century that individuals contributed to fields as diverse as poetry, politics, plant science, and physics (e.g., Johann Wolfgang Goethe, 1749–1832), the rapid increase in the amount of scientific information led to a diversification of disciplines. Currently, the vast amounts of scientific information gained over the last centuries are dispersed over thousands of journals, multiple languages, and, in the modern era, a variety of digital repositories for large-scale datasets (e.g., image stacks and omics datasets). Individual scientists are unable to keep track of all developments in their field, let alone to integrate information from different fields. While this problem can be alleviated partially by forming multidisciplinary collaborations, the extraction of relevant information from the exponentially increasing amounts of published data becomes ever more challenging. We suggest that the situation would improve if the publication of at least certain types of data occurred routinely by integration of that data into databases that structurally resemble the research objects in which the data was collected. Such databases represent the research object in its entirety; they are *holistic* in this sense. Once holistic models have become fully implemented, the function of scientific journals will have shifted. Their role as outlets for novel data will have vanished, and they will focus on providing venues for scientific discussion instead.

2. Current holistic models

In unrelated fields, the generation of holistic models has contributed to success stories. Attempts at predicting the weather probably are as old

as mankind, and accuracy was mainly based on individual experience in reading cloud patterns etc. Such experience-based predictions are strongly dependent on the location. Today, the accuracy for next-day forecasts approaches 100%, and that for five-day predictions is about 90% (<https://scijinks.gov/forecast-reliability/>) even at the most remote locations on the planet. This became possible due to the emergence of unifying models into which all available information could be fed. No meteorologist tasked with predicting the weather in one city for the next few days will start screening publications of tabulated weather data for this location over the last 100 years. Today, the fact that meteorologists predict local changes based on theoretical models that describe entire continents appears almost trivial. However, the transition to holistic atmospheric models was neither easy nor without resistance. The first mathematical models aiming at the prediction of a storm in 1950 provided a model in 1952, two years after that storm had occurred. A breakthrough was accomplished when meteorologists successfully predicted the behavior of the record-breaking ‘Blizzard of 1993’ in eastern north America. The prediction led to national alerts and states of emergency declared along the US east coast, saving countless lives (over 200 deaths occurred in the USA; <https://www.weather.gov/ilm/Superstorm93>). Only two decades earlier, influential meteorologists still had rejected attempts to model the atmosphere as a waste of time (for a more detailed account, we recommend the podcast <https://why.org/episodes/the-science-of-extreme-weather/>).

Wouldn't scientists in general benefit if similar, constantly updated models were available that would facilitate predictions of the responses of their experimental systems under a variety of conditions? Rather than feeding new data into such models, we still publish individual papers on related topics in a wide variety of outlets, forcing colleagues to search for each publication after they have noted it exists, archive it in their own analog or digital libraries, and finally process the information and integrate it into the virtual model in their own brains. In analogy to meteorological models, why do biologists not generate holistic models of the organisms of interest and deposit all (old and new) relevant data

^{*} Corresponding author. School of Biological Sciences Washington State University, PO Box 644236, Pullman, WA, 99164-4236, USA.

E-mail address: knoblauch@wsu.edu (M. Knoblauch).

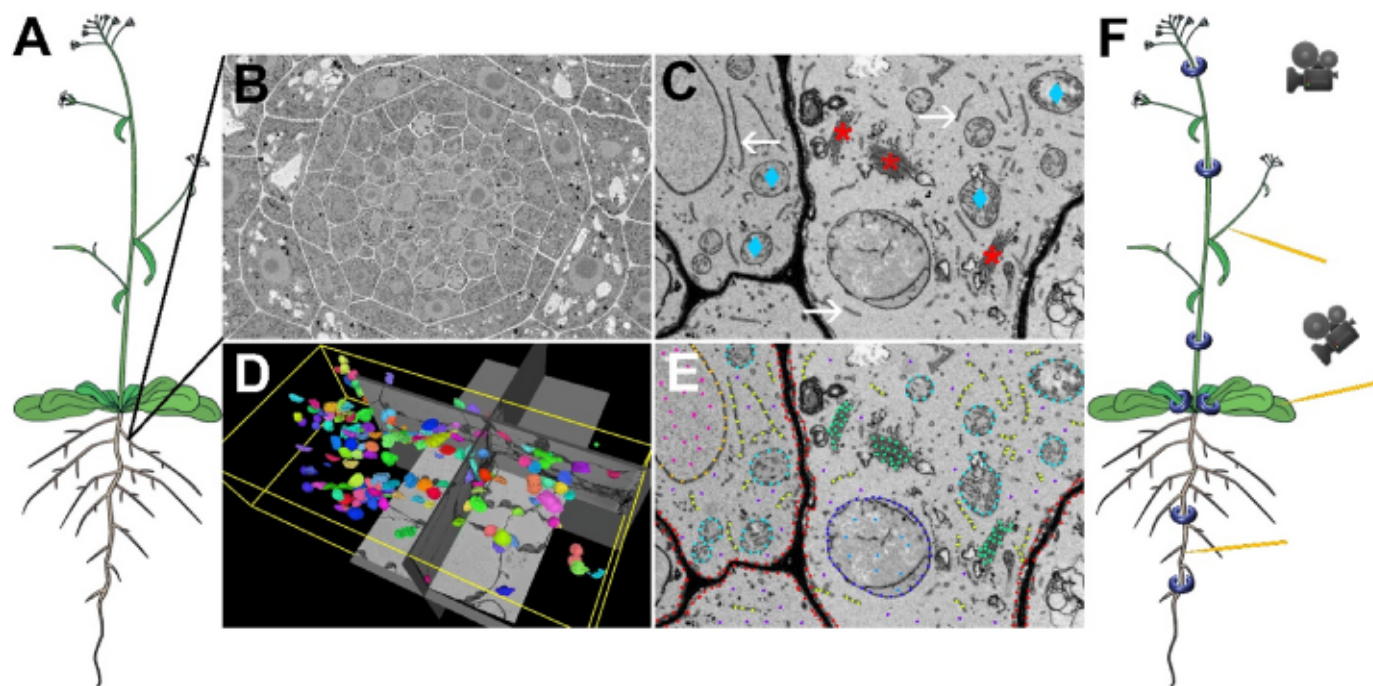


Fig. 1. Schematic representation of the road towards holistic organismic models.

A) Because the majority of experimental data originate from a few model organisms, it would make sense to initially focus on those systems to establish general principles. As plant scientists, we would suggest to work with *Arabidopsis*, but the same general approach would be applicable to *Drosophila*, *C. elegans*, and others. B, C) The first step would be the 3D reconstruction of the entire organism at subcellular level. For this aim, confocal and super resolution microscopy currently fall short as they have either insufficient resolution or a significant reduction of imaging quality towards deeper layers of tissues. Serial block face scanning electron microscopy currently would be the method of choice. It provides sufficient subcellular resolution (see the golgi, red asterisk; ER membranes, white arrows; and mitochondria, blue diamonds, in C) and due to the continuous removal of tissue layers in the imaging procedure, the resolution is consistent throughout the imaged block. Automated image acquisition by this method is established, but acquisition speed has its limitations. Therefore, other non-invasive imaging technologies such as x-ray microscopy may be more suitable in future once technical issues are solved.

D) Next, machine learning algorithms need to be developed and trained for automatic detection of cell and tissue types as well as subcellular structures. The image in D) shows an output file from the software Neuroglancer, highlighting the chloroplasts in leaf tissue. Appropriate algorithms are being developed, but significant improvements in detection accuracy are required.

E) Once an organism has been reconstructed and algorithms have generated a 3D grid with locations of cells, organelles, and other cell components, the model can be trained with physiological data that may initially include water permeability of the plasma membranes, conductivity through plasmodesmata, water fluxes within the xylem and phloem, cell turgor pressure, solute concentrations, and transporter localization (indicated by colored dots in E) etc. With increasing complexity temporal changes in transporter and protein expression will be included to reflect developmental changes or stress reactions (which may, of course, require decades).

F) Ultimately, the development of remote sensors which constantly feed the model with data depending on developmental state and environmental conditions will be critical for the success of holistic organismic models. Spectroscopic methods such as Raman and others appear potential methods of choice to remotely detect e.g. metabolic changes and (in the future) protein expression patterns, while developments such as X-ray microscopy, light sheet microscopy, or CT scanning may be improved to provide appropriate spatial and temporal resolution to follow developmental changes. In essence, algorithm and detector development will be critical for the generation of sophisticated holistic models.

in them? What if our essential data were deposited in databases designed in such a way that their internal structure as well as their user surfaces reflected the structures of the natural objects the data refer to?

Consider a printed list of all restaurants in all cities with their geographic coordinates tabulated, in comparison with a digital map in which all restaurants in all cities are marked. People in search of a restaurant clearly prefer the digital map over the table, as the success of software like MapQuest, Google Maps, Bing Maps, Maps.Me, etc., demonstrates. In contrast to tabulated data, an electronic map is a representation of the real thing, which makes working with it so much more intuitive. Couldn't we utilize such an electronic representation of the planet as a scaffold to structure the information contents of modern geosciences? Zooming into a location could lead us from the view of continents in a plate tectonics context to regional tectonic features, to the stratigraphic elements present at that place and their fossil content, and all the way to the crystallography and geochemistry of the local rocks. The 'map' would have to be 3D, to represent real geology underground and its inferred, eroded counterparts above ground level. Furthermore, the 'map' will have to be a time machine (4D), allowing users to move through geological history. On each zoom level at every

location, and at all geological times, all available pertinent information and its sources would be linked in, including print as well as non-printable media of all kind. Such a huge digital map could be further expanded by including information on water circulation, soil characteristics, vegetation types, etc., which would facilitate, for instance, the evaluation of predictions concerning the development of natural plant communities or of crop performance.

3. Holistic models in the near future

Databases and user interfaces that are structured to mimic the real object could be designed in many disciplines. As plant physiologists, our interest focuses on an electronic plant model. We envision 3D reconstructions of plants at subcellular resolution. Recent developments in computer and visualization technologies have enabled the collection of large imaging datasets and resulted in the introduction of confocal, super-resolution, and serial-block-face-imaging methodologies that allow for automated, large-scale image acquisition at resolutions down to 10 nm. A small plant like *Arabidopsis thaliana* reconstructed at 10 nm resolution will already require petabytes of storage space. The temporal

aspect of the plant's representation would be achieved by combining models of plants at different developmental stages which would add accordingly. We conclude that while the reconstruction of whole organisms by these technologies is possible in principle, significant further improvements doubtlessly are necessary to turn the procedures into large-scale routine tasks. For instance, machine learning algorithms for the automatic identification of cellular and subcellular structures including organelles, cell walls, plasmodesmata, and membranes as well as for the integration of transporters and other relevant proteins, will be required to accelerate the integration of physiological meaning into the structural model – which leads us to an important difference between, on one hand, maps that merely link all existing, relevant information to the appropriate structures, and prognostic models that help predicting the behavior of complex systems like our weather on the other hand.

Today, publicly accessible geographic models integrate geographic structure with information about weather conditions, traffic rules including speed limits and one-way traffic, construction activities, and cultural features such as holidays and large public events, opening hours of shops and gastronomy, to predict traffic conditions from which recommendations for traveling users are derived. Similarly, a holistic model of, for instance, an Arabidopsis could be designed to predict and help explaining transport processes in xylem and phloem, both on the local and whole-organism level and for a variety of outer conditions. Transport to and from these long-distance pathways, such as water fluxes toward the central cylinder in roots or sugar movement toward the phloem in photo-assimilating source organs, could be included. To this end, quantitative information on functional parameters will have to be integrated in the model. On the cellular level, for example, combining 3D structural data with diffusional membrane conductances for certain molecules, all relevant membrane transporter activities, and plasmodesma densities and conductivities will allow to model cellular exchange rates of the molecules and compare them to experimental results. This could show whether crucial components of the overall transport process are missing from the model. Refinement of model predictions based on experimental data will lead to increased accuracy until a satisfying degree of correspondence between model and experiment is reached. At that stage, model predictions will have attained a potential to guide discoveries and become the “1993 Blizzard of Plant Science”. A structurally-functionally integrated holistic model could help predicting phenotypes and organismal behavior in dependence of environmental factors. Explorative phenotype studies, for example in the context of crop improvement, could be performed *in silico* rather than through year-long studies on transgenic lines. While interesting hypotheses

obviously would have to be tested and confirmed in reality, the overall process could be drastically shortened.

4. Long-term development of dynamic holistic models

Ultimately we will become able to add automatic remote sensing to lab-based experimental approaches, partially replacing the latter. The dynamic character of atmospheric models is based on optimized sensors that are distributed throughout the globe and send a set of critical data that constantly feed the model to allow dynamic adjustments. Similarly, most of us serve as remote sensors to feed the dynamic model via our cell phones for traffic prediction. The development of sensors providing dynamic data every second on physiology, metabolite changes, expression patterns, water fluxes, 3D anatomy etc. will allow plant developmental modeling in dependence of biotic or abiotic stressors. Remote sensing will also reshape the undoubtedly difficult task of data quality control that currently develops into a major problem in various disciplines. Peer reviewing as it is known to date would become obsolete. Journals would transform into discussion platforms without publication of original data. Experiments would focus on the testing of model predictions.

Phenomics tools already employ remote sensors to monitor photosynthesis and other parameters in plants. Smart watches track basic physiological functions of hundreds of millions of humans. There is no doubt that these and similar technologies will improve significantly in the coming decades, which will facilitate automated model feeding.

Changing a scientific communication system that has worked well for generations predictably will face skepticism. However, we anticipate that current developments will lead to holistic models in one or the other form. We believe that such models would massively simplify the intuitive access for researchers to data, provide guidance in understanding organismal functions, and drastically improve data accuracy and meaningful output. As the tools to initiate such efforts are now available, we suggest to get started.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We thank Jan Knoblauch for providing Fig. 1D. MK acknowledges funding by NSF PGRP 1940827.