Open Access

How to MyST, without being mystified 🧙

Evolve your markdown documents into structured data

Rowan Cockett[1,2] [1]Executable Books, [2]Curvenote

Abstract

We introduce, a set of open-source, community-driven …

1. MyST 사용법, 미스터리 없이!🧙

마크다운 문서와 노트북을 구조화된 데이터로 변환하는 튜토리얼

Authors: Rowan Cockett[1,2]
Affiliations: [1]Executable Books, [2] Curvenote
License: CC-BY

Abstract

저희는 블로그, 온라인 서적, 과학 논문, 보고서 및 저널 기사를 지원하는 강력한 저작 프레임워크를 포함하여 과학 커뮤니케이션을 위해 설계된 MyST 마크다운 (myst.tools)용 오픈 소스, 커뮤니티 주도형 도구 세트를 소개합니다.

1.a. Background

오늘날의 과학 커뮤니케이션은 인쇄 문서와 유료 콘텐츠 액세스를 중심으로 설계되어 있습니다. 지난 10년간 오픈 사이언스 운동은 사전 인쇄 서비스 및 데이터 아카이브의 사용을 가속화하여 과학 콘텐츠의 접근성을 크게 향상시켰습니다. 그러나 이러한 시스템은 학술 문헌의 논문 중심 모델보다 훨씬 더 많은 것을 포함하는 현대 과학적 결과물을 전달하도록 설계되지 않았습니다.

> 우리는 과학 지식을 공유하고 소통하는 방식이 인쇄 기반 출판의 현상 유지와 종이의 모든 한계를 넘어 진화해야 한다고 믿습니다.

The communication and collaboration tools that we are building in the Project Jupyter are built to follow the FORCE11 recommendationsBourne et al. (2012). Specifically:

1. rethink the unit and form of scholarly publication;
2. develop tools and technologies to better support the scholarly lifecycle; and
3. add data, software, and workflows as first-class research objects.

Published Jul 08, 2025

By bringing professional, high-quality tools for science communication into the research lifecycle, we believe we can improve the collection and preservation of scholarly metadata (citations, cross-references, annotations, etc.) as well as open up new ways to communicate science with interactive figures & equations, computation, and reactivity.

The tools that are being built by the Project Jupyter are focused on introducing a new Markup language, MyST (Markedly Structured Text), that works seamlessly with the Jupyter community to enhance and promote a new path to document creation and publishing for next-generation scientific textbooks, blogs, and lectures. Our team is currently supported by the Sloan Foundation, (Grant #9231).

MyST enables rich content generation and is a powerful format for scientific and technical communication. JupyterBook uses MyST and has broad adoption in publishing tutorials and educational content focused around Jupyter Notebooks.

> The components behind Jupyter Book are downloaded 30,000 times a day, with 750K downloads last month.

The current toolchain used by JupyterBook is based on Sphinx, which is an open-source documentation system used in many software projects, especially in the Python ecosystem. `mystjs` is a similar tool to Sphinx, however, designed specifically for scientific communication. In addition to building websites, `mystjs` can also help you create scientific PDFs, Microsoft Word documents, and JATS XML (used in scientific publishing).

`mystjs` uses existing, modern web-frameworks in place of the Sphinx build system. These tools come out-of-the-box with prefetching for faster navigation, smaller network payloads through modern web-bundlers, image optimization, partial-page refresh through single-page application. Many of these features, performance and accessibility improvements are difficult, if not impossible, to create inside of the Sphinx build system.

In 2022, the Executable Books team started work to document the specification behind the markup language, called myst-spec, this work has enabled other tools and implementations in the scientific ecosystem to build on MyST (e.g. scientific authoring tools, and documentation systems).

The `mystjs` ecosystem was developed as a collaboration between Curvenote, 2i2c and the ExecutableBooks team. The initial version of `mystjs` was originally release by Curvenote as the Curvenote CLI under the MIT license, and transferred to the ExecutableBooks team in October 2022. The goal of the project is to enable the same rich content and authoring experiences that Sphinx allows for software documentation, with a focus on web-first technologies (Javascript), interactivity, accessibility, scientific references (e.g. DOIs and other persistent IDs), professional PDF outputs, and JATS XML documents for scientific archiving.

1.b. MyST Project

In this paper we introduce `mystjs`, which allows the popular MyST Markdown syntax to be run directly in a web browser, opening up new workflows for components to be used in web-based editors, directly in Jupyter and in JupyterLite. The libraries work with current MyST Markdown documents/projects and can export to LaTeX/PDF, Microsoft Word and JATS as well as multiple website templates using a modern React-based renderer. There are currently over 400 scientific journals that are supported through templates, with new LaTeX templates that can be added easily using a Jinja-based templating package, called jtex.

In our paper we will give an overview of the MyST ecosystem, how to use MyST tools in conjunction with existing Jupyter Notebooks, markdown documents, and JupyterBooks to create professional PDFs and interactive websites, books, blogs and scientific articles. We give special attention to the additions around structured data, standards in publishing (e.g. efforts in representing Notebooks as JATS XML), rich frontmatter and bringing cross-references and persistent IDs to life with interactive hover-tooltips (ORCID, RoR, RRIDs, DOIs, intersphinx, wikipedia, JATS, GitHub code, and more!). This rich metadata and structured content can be used directly to improve science communication both through self-publishing books, blogs, and lab websites — as well as journals that incorporate Jupyter Notebooks.

1.c. Features of MyST

MyST is focused on scientific writing, and ensuring that citations are first class both for writing and for reading (see Figure 1).

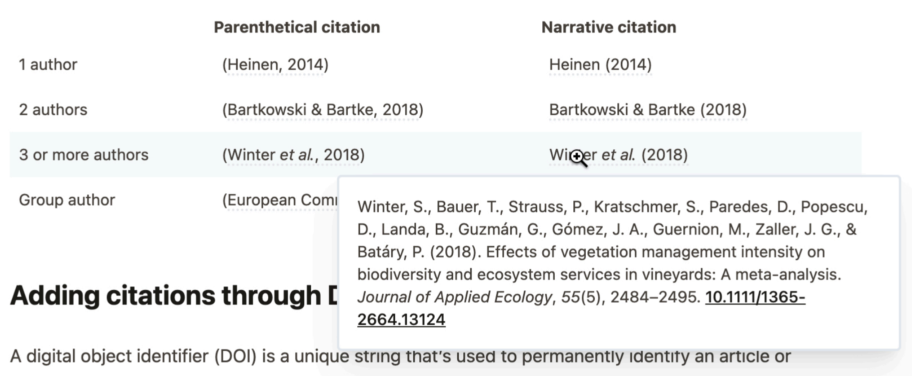
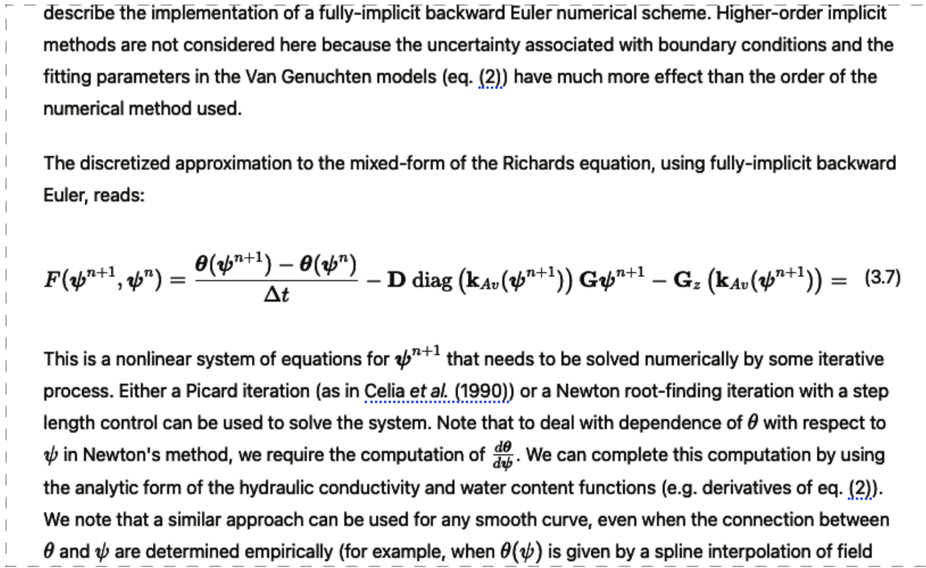


Figure 1: Citations are rendered with a popup directly inline.

MyST aims to show as much information in context as possible, for example, Figure 2 shows a reading experience for a referenced equation: you can immediately **click on the reference**, see the equation, all without loosing any context – ultimately saving you time. Head et al. (2021) found that these ideas both improved the overall reading experience of articles as well as allowed researchers to answer questions about an article **26% faster** when compared to a traditional PDF!



describe the implementation of a fully-implicit backward Euler numerical scheme. Higher-order implicit methods are not considered here because the uncertainty associated with boundary conditions and the fitting parameters in the Van Genuchten models (eq. (2)) have much more effect than the order of the numerical method used.

The discretized approximation to the mixed-form of the Richards equation, using fully-implicit backward Euler, reads:

$$F(\boldsymbol{\psi}^{n+1}, \boldsymbol{\psi}^{n}) = \frac{\boldsymbol{\theta}(\boldsymbol{\psi}^{n+1}) - \boldsymbol{\theta}(\boldsymbol{\psi}^{n})}{\Delta t} - \mathbf{D}\ \mathrm{diag}\left(\mathbf{k}_{Av}(\boldsymbol{\psi}^{n+1})\right) \mathbf{G}\boldsymbol{\psi}^{n+1} - \mathbf{G}_z \left(\mathbf{k}_{Av}(\boldsymbol{\psi}^{n+1})\right) = \quad (3.7)$$

This is a nonlinear system of equations for $\boldsymbol{\psi}^{n+1}$ that needs to be solved numerically by some iterative process. Either a Picard iteration (as in Celia *et al.* (1990)) or a Newton root-finding iteration with a step length control can be used to solve the system. Note that to deal with dependence of θ with respect to ψ in Newton's method, we require the computation of $\frac{d\theta}{d\psi}$. We can complete this computation by using the analytic form of the hydraulic conductivity and water content functions (e.g. derivatives of eq. (2)). We note that a similar approach can be used for any smooth curve, even when the connection between θ and ψ are determined empirically (for example, when $\theta(\psi)$ is given by a spline interpolation of field

Figure 2: In context cross-references improve the reading experience.

One of the important underlying goals of practicing reproducibility, sharing more of the methods and data behind a scientific work so that other researchers can both verify as well as build upon your findings. One of the exciting ways to pull for reproducibility is to make documents directly linked to data and computation! In Figure 3, we are showing outputs from a Jupyter Notebook directly part of the published scientific narrative.

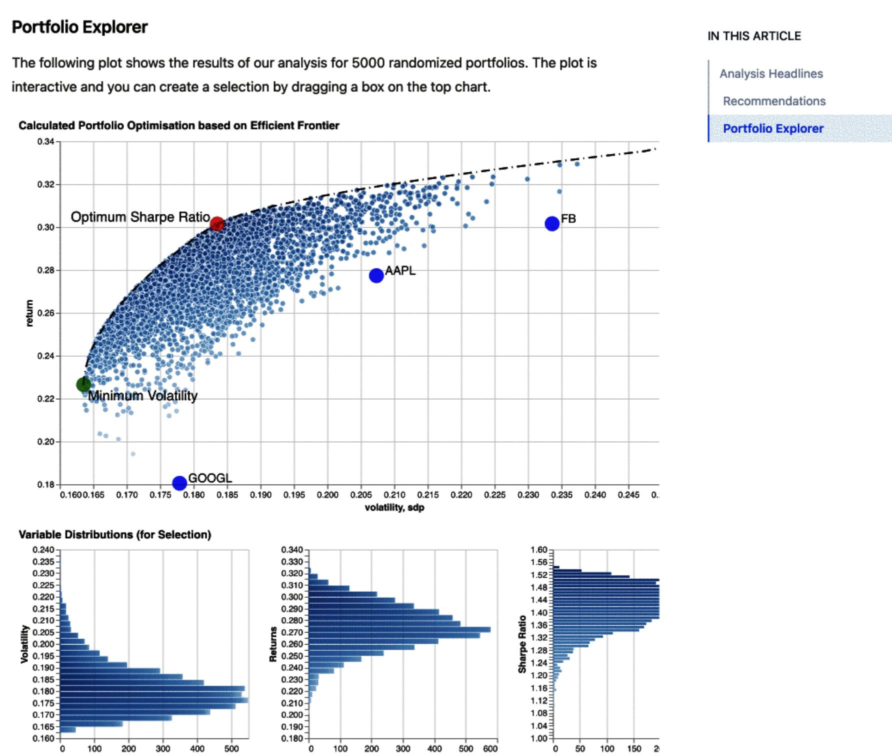


Figure 3: Embedding data, interactivity and computation into a MyST article.

To drive all of these features, the contents of a MyST document needs to be well defined. This is critical for powering interactive hovers, linked citations, and compatibility with scientific publishing standards like the Journal Article Metadata Tag Suite (JATS). We have an emerging specification for MyST, `myst-spec`, that aims to capture this information and transform it between many different formats, like PDF, Word, JSON, and JATS XML (Figure 4). This specification is arrived at through a community-centric MyST Enhancement Proposal (MEP) process.

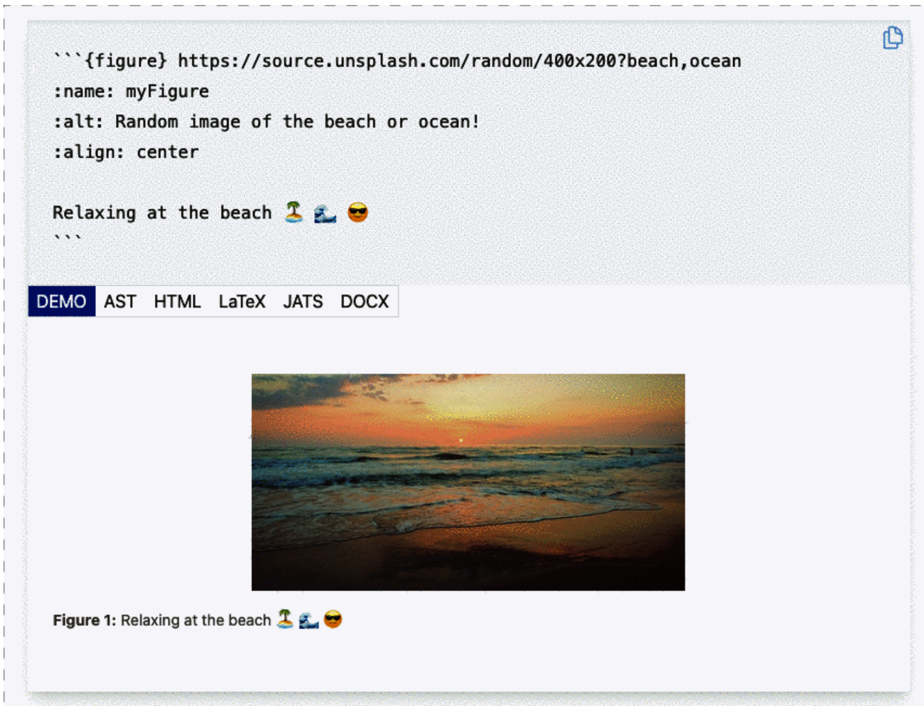
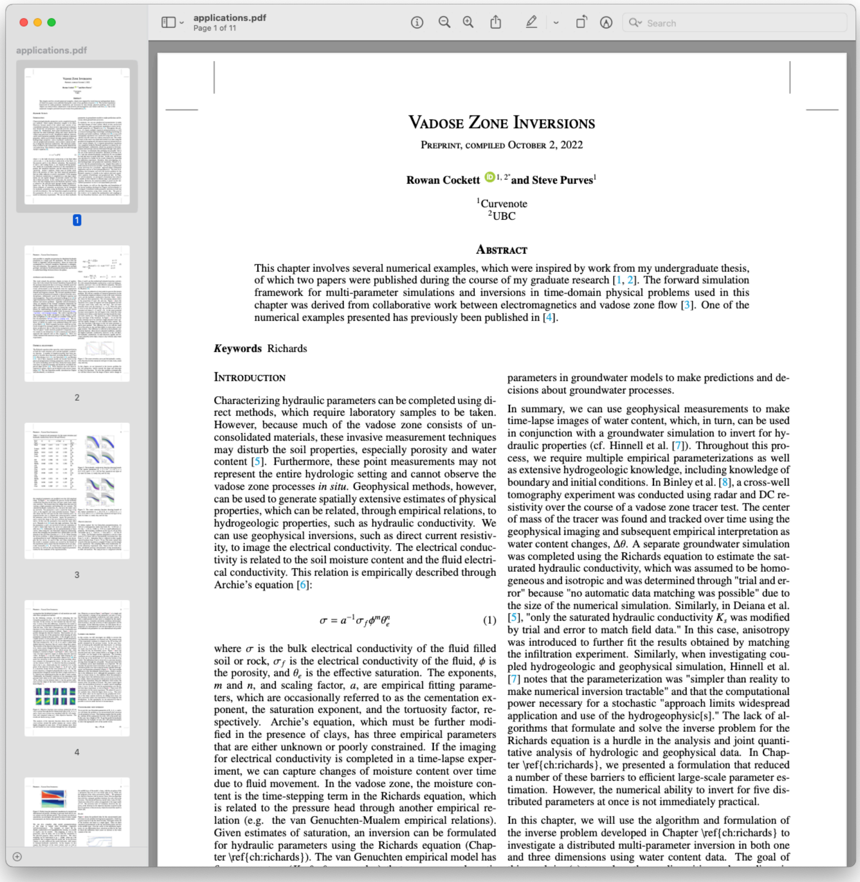


Figure 4: The data behind MyST is **structured**, which means we can transform it into many different document types and use it to power all sorts of exciting features!

One of the common forms of scientific communication today is through PDF documents. MyST has excellent support for creating PDF documents, using a data-driven templating library called `jtex`. The document in Figure 5 was created using MyST!



applications.pdf
Page 1 of 11

Vadose Zone Inversions

Preprint, compiled October 2, 2022

Rowan Cockett [1, 2*] and Steve Purves [1]

[1] Curvenote
[2] UBC

Abstract

This chapter involves several numerical examples, which were inspired by work from my undergraduate thesis, of which two papers were published during the course of my graduate research [1, 2]. The forward simulation framework for multi-parameter simulations and inversions in time-domain physical problems used in this chapter was derived from collaborative work between electromagnetics and vadose zone flow [3]. One of the numerical examples presented has previously been published in [4].

Keywords Richards

Introduction

Characterizing hydraulic parameters can be completed using direct methods, which require laboratory samples to be taken. However, because much of the vadose zone consists of unconsolidated materials, these invasive measurement techniques may disturb the soil properties, especially porosity and water content [5]. Furthermore, these point measurements may not represent the entire hydrologic setting and cannot observe the vadose zone processes *in situ*. Geophysical methods, however, can be used to generate spatially extensive estimates of physical properties, which can be related, through empirical relations, to hydrogeologic properties, such as hydraulic conductivity. We can use geophysical inversions, such as direct current resistivity, to image the electrical conductivity. The electrical conductivity is related to the soil moisture content and the fluid electrical conductivity. This relation is empirically described through Archie's equation [6]:

$$\sigma = a^{-1}\sigma_f \phi^m \theta_e^n \tag{1}$$

where σ is the bulk electrical conductivity of the fluid filled soil or rock, σ_f is the electrical conductivity of the fluid, ϕ is the porosity, and θ_e is the effective saturation. The exponents, m and n, and scaling factor, a, are empirical fitting parameters, which are occasionally referred to as the cementation exponent, the saturation exponent, and the tortuosity factor, respectively. Archie's equation, which must be further modified in the presence of clays, has three empirical parameters that are either unknown or poorly constrained. If the imaging for electrical conductivity is completed in a time-lapse experiment, we can capture changes of moisture content over time due to fluid movement. In the vadose zone, the moisture content is the time-stepping term in the Richards equation, which is related to the pressure head through another empirical relation (e.g. the van Genuchten-Mualem empirical relations). Given estimates of saturation, an inversion can be formulated for hydraulic parameters using the Richards equation (Chapter \ref{ch:richards}). The van Genuchten empirical model has

parameters in groundwater models to make predictions and decisions about groundwater processes.

In summary, we can use geophysical measurements to make time-lapse images of water content, which, in turn, can be used in conjunction with a groundwater simulation to invert for hydraulic properties (cf. Hinnell et al. [7]). Throughout this process, we require multiple empirical parameterizations as well as extensive hydrogeologic knowledge, including knowledge of boundary and initial conditions. In Binley et al. [8], a cross-well tomography experiment was conducted using radar and DC resistivity over the course of a vadose zone tracer test. The center of mass of the tracer was found and tracked over time using the geophysical imaging and subsequent empirical interpretation as water content changes, $\Delta\theta$. A separate groundwater simulation was completed using the Richards equation to estimate the saturated hydraulic conductivity, which was assumed to be homogeneous and isotropic and was determined through "trial and error" because "no automatic data matching was possible" due to the size of the numerical simulation. Similarly, in Deiana et al. [5], "only the saturated hydraulic conductivity K_s was modified by trial and error to match field data." In this case, anisotropy was introduced to further fit the results obtained by matching the infiltration experiment. Similarly, when investigating coupled hydrogeologic and geophysical simulation, Hinnell et al. [7] notes that the parameterization was "simpler than reality to make numerical inversion tractable" and that the computational power necessary for a stochastic "approach limits widespread application and use of the hydrogeophysic[s]." The lack of algorithms that formulate and solve the inverse problem for the Richards equation is a hurdle in the analysis and joint quantitative analysis of hydrologic and geophysical data. In Chapter \ref{ch:richards}, we presented a formulation that reduced a number of these barriers to efficient large-scale parameter estimation. However, the numerical ability to invert for five distributed parameters at once is not immediately practical.

In this chapter, we will use the algorithm and formulation of the inverse problem developed in Chapter \ref{ch:richards} to investigate a distributed multi-parameter inversion in both one and three dimensions using water content data. The goal of

Figure 5: A PDF rendering through MyST.

1.d. *Conclusion*

There are many opportunities to improve open-science communication, to make it more interactive, accessible, more reproducible, and both produce and use structured data throughout the research-writing process. The `mystjs` ecosystem of tools is designed with structured data at its core. We would love if you gave it a try – learn to get started at https://myst.tools.

1.e. *References*

Bourne, Philip E., Clark, Timothy W., Dale, Robert, De Waard, Anita, Herman, Ivan, Hovy, Eduard H., Shotton, David. (2012)"Improving The Future of Research Communications and e-Scholarship". FORCE11. doi:10.4230/DAGMAN.1.1.41

Head, A., Lo, K., Kang, D., Fok, R., Skjonsberg, S., Weld, D. S., & Hearst, M. A. (2021, May). Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 10.1145/3411764.3445648

References

Bourne, P. E., Clark, T. W., Dale, R., Waard, A. de, Herman, I., Hovy, E. H., & Shotton, D. (2012). *Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331)*. https://doi.org/10.4230/DAGMAN.1.1.41

Head, A., Lo, K., Kang, D., Fok, R., Skjonsberg, S., Weld, D. S., & Hearst, M. A. (2021). Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3411764.3445648