

IDC workshop at the 2023 Gran Canaria Project Week

Logistics: Wednesday, February 1, 2023 4pm-6pm Las Palmas time (11am-1pm US Eastern time) - in person and via zoom:

<https://etsmtl.zoom.us/j/86211702920?pwd=TEl0ZTFDam90WVN5bjZhR05kNVRVZz09>

Vote in the poll following this link: <https://app.sli.do/event/8DEbd4iq2ByirPgFDtSPR8>

NCI Imaging Data Commons (IDC) is a cloud-based repository of publicly available cancer imaging data co-located with the analysis and exploration tools and resources. IDC is a node within the broader NCI [Cancer Research Data Commons \(CRDC\)](#) infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data.

IDC Portal is the interactive interface that allows exploring data available in IDC using a small subset of metadata attributes accompanying IDC data, visualize radiology and microscopy images and annotations, save cohorts (subsets of data) under user account based on the available metadata filters.

The real power of IDC comes, however, from programmatic interfaces available to work with IDC data. Most of the capabilities available through those interfaces and APIs are not available in the portal. As few examples

- download of the image files available in IDC (portal allows you to download a manifest, but download of files referenced from the manifest is currently not easily doable)
- flexible definition of the selection filters at the level of series, studies or patients (portal exposes a very tiny set of metadata attributes, does not have the flexibility in defining filters that is needed for analysis tasks)
- portal does not expose clinical metadata available for imaging collections

In this tutorial/workshop you will:

- Learn what is IDC and how it can be useful to you
- See some examples of analysis use cases that utilize the cloud to evaluate existing AI tools on public data and utilize state of the art approaches to annotate data in IDC
 - including discussions about costs and trade-offs using different cloud components for accomplishing this task
- Understand the resources available to start using IDC programmatically
- Become introduced to the projects related to IDC that
 - facilitate sharing of analysis tools that are designed to be easily applicable to work with DICOM
 - enable scalable application of analysis tools to large data cohorts

Planned agenda - Slides are now available publicly!

1. **[4-4:10pm] Introduction to NCI Imaging Data Commons and outline of the workshop** - 10 min (Andrey Fedorov)
 - a. Slides  PW38 IDC intro
2. **[4:10pm-4:30pm] Use case: Lung CT analysis with Body Part Regression and nnU-Net + QA** - 20 min (Deepa Krishnaswamy)

- a. Slides  PW_Deepa_and_Cosmin
 - b. Analysis results shared on Zenodo <https://zenodo.org/record/7539035> (release in IDC underway)
 - c. QA
3. **[4:30pm-4:50pm] Use case: Prostate MRI analysis with nnU-Net and MONAI + QA** - 20 min (Cosmin Ciausiu)
 - a. Slides  PW_Deepa_and_Cosmin
4. **[4:50pm-5pm] Discussion** - 10 min
5. **[5pm-5:15pm] How to get started with IDC:** 15 min (Andrey Fedorov)
 - a. Overview of the [IDC Portal](#)
 - b. Overview of the [hands-on colab notebooks tutorial](#)
 - c. FYI: working with clinical data - [intro notebook](#)
6. **[5:15pm-5:30pm] Infrastructure for curation and distribution of AI models:** MHub ([ModelHub.AI](#)) - 15 min (Dennis Bontempi and Leonard Nuernberg)
 - a. Slides  MHub @ PW (IDC)
7. **[5:30pm-5:40pm] Applying AI models at scale:** [Seven Bridges Genomics](#) and [Terra](#) - 10 min (Vamsi Thiriveedhi)
 - a. Slides  SBG and Terra
8. **5:40pm-5:50pm Discussion** - 10 min
9. **5:50pm-6pm Conclusions** +- 10 min
 - a. Slides  PW38 IDC intro
 - b. Rehash the overall workflow
 - c. Plans for the future development
 - d. Surprise announcement about a relevant funding opportunity: <https://sam.gov/opp/c8e97bbc54944eda92645e1629df8228/view>
 - e. Discussion + Feedback collection
10. **6pm-7pm Kaapana presentation** +- 60 min (Klaus Kades)
 - a. Slices  Na-MIC Kaapana tutorial
 - b. Presentation of Kaapana
 - c. How to develop with Kaapana

Questions / Comments

Please add your questions or comments below. If you did anything with IDC and want to share your experience, please add a note as well.

•

Notes of questions from audience during 4:50-5 pm Discussion:

- Question about costs:
 - Can run queries without paying anything
 - If you have a billing account set up with your account you should set up billing alerts
 - In IDC you can apply for cloud credits easily - \$300
- Question about query:
 - How do we indicate which volumes are valid? Adding column to dicom metadata to indicate this?
- XNat vs BigQuery
- Question about NLST query - where is the documentation, how do we know which columns, hierarchical format
 - You can refer to the getting started notebook: https://github.com/ImagingDataCommons/IDC-Examples/blob/master/notebooks/getting_started.ipynb and notebooks here: https://github.com/ImagingDataCommons/IDC-Examples/tree/master/notebooks/getting_started

Notes of questions from audience during 5:40-5:50 Discussion:

- **Q - Have we thought about adding a layer on top of SQL BigQuery to make it simpler?**
 - **A - does not want to add another layer, might be limitations**
 - **A - providing the result of query in IDC would help**
- **Q - what about Django?**
- **Q - Did we think about having an ORM? (object relational mapping)**
 - **A - not sure about effort yet**
- **Q - Query - difference is cohort is static**
 - **Get content of table**
- **Q - how to contribute to these datasets?**
 - **A - not a straightforward way, but depends on what type of dataset. If the dataset is already in IDC, and we don't have annotations – we have a good precedent. But if you have a dataset from your institution etc, need to de-identify, IDC will not do that. You can try TCIA instead. Or what about slide microscopy, need converting to DICOM, deidentify, etc**