WGBH - DCAMM SCRUM Report 1	
Contact	Github accounts: elisa3lopes, rlee99, murtio, carmen-araujo, jenajjedu Email addresses: ● Jena Jordahl jenajj@bu.edu ● Elisa Cordeiro Lopes elisacl@bu.edu ● Richard Lee rlee99@bu.edu ● Murtadha Bahrani murtadha@bu.edu ● Carmen Sabrina Araujo sabrinaa@bu.edu
What have I worked on?	● Built the initial parser, still under development. ● Talked with professor and established smaller steps for answering questions on the data ● Setting up Grobid on SCC. Grobid is used to parse PDFs into XML, we are trying to configure parsing from tables. ● We tried 7 different ways to parse the PDF (real time, tabula software, tabula library, Grobid, pypdf2, managing data after tabula, transforming data back to pdf)
Have I talked to the client recently? When are we meeting with them next?	● We talked with the client on Wed 3/4/21 11:30-12:00pm. ● The client understood our stage of development and checked we are on the correct path ● Third meeting with the client on Thurs March 11th
What will I be working on next?	● Next week we are going to add tests for test driven development: ● Three levels of test: Companies, Trades, and Hours reported by the companies per trade. ● Contact SCC to request computational resources. ● Experiment with Grobid to see if we can use it for parsing PDF tables.

	● Take the data and reverse engineer by how the PDF data was created, printed by using row types.
Have I run into any issues? Do I need help?	● Pandas dataframes and csv cells still merged. How should we deal with this problem? Should we keep this strategy or change? ● CSV not congruous throughout. Will print csv and find out its structure.