Using Large Scale Video Data to Study Organizational Processes

Marshall Scott Poole
Natalie J. Lambert
Sandeep P. Satheesan
Amit Das
Sujeeth Bharadwaj
Gabriel Merrin
Mark Hasegawa-Johnson
Alex Yahja
University of Illinois Urbana-Champaign

Noshir Contractor
Northwestern University

Kenton McHenry
Rob Kooper
University of Illinois Urbana-Champaign

Melissa Dobosh
University of Northern Iowa

Dorothy Espelage
Margaret Fleck
David Forsyth
University of Illinois Urbana-Champaign

Feniosky Pena-Mora
Columbia University

Correspondence should be addressed to Marshall Scott Poole, Department of Communication, University of Illinois Urbana-Champaign, 3001 Lincoln Hall, 702 S. Wright St, Urbana, IL 61801, mspoole@illinois.edu

.

Using Large-Scale Video Data to Study Organizational Processes

Abstract

Video is an ideal tool for the study of organizational processes. For even moderate-sized units, however, organizational processes typically unfold over significant periods of time—hours to months—and involve a large and often changing set of participants. Current approaches to utilizing video to study processes are tailored to following individuals or relatively small, well-bounded groups, and the sheer amount and complexity of video data required to study larger-scale processes presents unique challenges that current approaches are ill-suited to address. This paper describes an integrated system for observational research on larger-scale organizational processes using video, supplemented by other forms of data. It discusses requirements for such a system, a basic design, some of the tools available or in development to implement the system, and outstanding challenges. While integrated systems such as that envisioned in this paper are still in the future, there has been noticeable progress. (146 words)

Recent currents in organizational scholarship have conceptualized organizations as processes, which "prioritizes activity over outcome, change over persistence, novelty over continuity, expression over determination, and becoming over being. Accordingly flux and transformation as well as creativity, disruption, and indeterminism are key themes within a process worldview" (Schultz, Maguire, Langley & Tsoukas, 2012).

Organizational research has long taken processes into account (e.g., Mintzberg, 1979), but until the turn of the 21st century process was largely subordinated to structure and was treated as though it unfolded within the constraints of organizational structure. Recent scholarship has sought to release process from its structural cage. A wide range of organizational phenomena have been studied using a processual lens, from microlevel phenomena such as identity construction (Pratt, 2012) to innovation (Van de Ven, Angle, & Poole, 2000) and sensemaking (Hernes & Maitlis, 2011) at the organization level, to interorganizational networks (Monge, Heiss & Margolin, 2008) at the macro-level. Theoretical sophistication in the study of processes has expanded rapidly during this period (Van de Ven & Poole, 1995; Langley & Tsoukas, 2010).

Organizational processes typically unfold over significant periods of time—hours to months—and involve a large and often changing set of participants. A variety of methods have been employed to gather data suitable for process research (e.g., Langley, 1999; Poole, Van de Ven, Dooley & Holmes, 2000), including ethnography and participant observation, collection of historical and transactional records, and longitudinal surveys. These have yielded valuable insights, but each type of data has limitations. Ethnography and participant observation produce synoptic accounts of processes that may elude important details. Historical and transactional records are limited to what participants or those designing the records deemed important. Surveys produce relatively thin data that is limited to concepts researchers deemed important.

These methods are often combined in process research in order to overcome the limitations due to any single source of data.

A particularly valuable and underutilized approach is direct observation of processes enabled by the capture of video, audio, and digital data. Observational recordings preserve a record of the process that can be analyzed multiple times and revisited at the researcher's discretion. Direct observation is an exceptionally valuable tool for process research, because it facilitates systematic analysis, retrospective evaluation of conjectures and interpretations, and layered studies that illuminate multiple dimensions of the process. Observational data can also be used to discover patterns in organizational processes that may not have been foreseen by theory. Video and audio have long been used in the study of relatively small, well-bounded social units, such as dyads and groups meeting in a room around a table. However, it is another matter to collect and to analyze video (audio and digital traces) in larger-scale organizations; doing so involves much greater complexity. Consider the three examples shown in Table 1, which represent organizations that are often regarded as stable, but which are actually quite dynamic.

To study anything but the most circumscribed processes in these settings, researchers must capture video of a large number of participants working in complex organizational structures that may change over time. The boundaries between subunits and between the organization and its environment are fuzzy, and participants may belong to multiple units and may serve different roles at different times and places, making identification of members on the basis of role or unit difficult. Personnel are distributed spatially and temporally and they are likely to move about and change locations during the period of observation, showing up on different cameras and on different audio recordings over time. Membership in the organization

A nursing floor in a hospital runs two 12 hour shifts a day with two different sets of nurses (each working three or four 12 hour days/week) for each shift. Nurses during each shift work in flexible teams

to coordinate care for patients; these teams reconfigure depending on the specific mix of patients and their needs. Nurses working the two shifts must coordinate patient care at the handoff time and nurses in one set must coordinate with those in the other when a new set comes on. Nurses from the floor also serve on several committees, including: five who comprise a quality improvement committee whose members also include a couple of physicians and a facilitator from human resources; a patient safety committee of five nurses; and two who serve as liaisons to the General Nursing Management Committee which sets general policy for all floors in the hospital. This hospital floor can be viewed as one large team, as four teams comprised of nurses in the same shift and set, or as seven teams, if the three additional committees are added. The nurses also have to coordinate on an *ad hoc* basis with physicians and physical therapists to deliver care, improvising small temporary teams "on the fly."
An emergency response incident command team convenes in the county response center to coordinate response to a train derailment that has resulted in a chemical spill and fire. The incident command is comprised of two city fire chiefs, a county fire chief, city and county police chiefs, a city manager, and the director of the local Red Cross. The team must gather data from multiple sources, make sense of the situation, and plan a response, including assignment of personnel to teams, task assignment, supervision of responder teams, updating plans as new information comes in, and communication with the public about the incident. As the group works, its members will reconfigure into subgroups that deal with different aspects of the situation and work out elements of response. Some members then go out into the field to supervise teams of responders, effectively creating multi-team complexes. These field teams will operate for a twelve hour shift, at which time the incident management team will reconvene to re-evaluate its plan and to pass off responsibility to a new incident management team, which will continue for the next twelve hour shift, and so on until the incident is declared over.
An elementary school holds recess for several classes. Students first convene in a common room, supervised by teachers, and then line up to go out for recess. Once everyone is quiet, one of the teachers signals that they can go to the playground. The children scatter across the playground and teachers space themselves out so they can monitor the students. Clusters of students form and break up throughout the recess, often around spaces such as a basketball court or pieces of equipment such as a jungle gym. In addition to play, behavior such as bullying and taunting occur, and teachers attempt to minimize this when they see it. Teachers also use the recess time to discuss other work-related matters and to socialize with one another. When the bell rings, students move back in and sort into their classes, with some residual impact from the recess such as hurt feelings, bonds made, and general activation carrying over into the classrooms.

Table 1. Three example organizational venues for the study of processes

over time, not only as a function of turnover, but because changes in activities require changes in membership.

To record the action in one of the organizational units or organizations just described requires capture of multiple video and audio streams, in addition to documents and digital data. For a 40-person emergency response organization moving over the scene of a disaster simulation, this might require 50 video cameras placed about the lot and forty audio recordings from individual participants running throughout a six hour simulation. The result is literally

hundreds of hours of video and audio recordings and gigabytes of digital data that must be analyzed.

The traditional approach to observational analysis in which researchers watch and annotate all the video "by hand" is prohibitively expensive in terms of time and resources in this case. It is, however, important to find ways to do so, because direct observational research offers an important window into organizational processes, a valuable additional viewpoint to compensate for the inevitable blindspots in any single approach.

This paper is concerned with one possible solution to the problems inherent in large-scale observational studies. It focuses on opportunities and challenges in utilizing computational video, audio, and text analytics and data integration technologies to manage data and analysis of large video corpora. It also advances a general design for an analytical environment that cuts the task of observational research on large-scale organizational processes down to manageable proportions. Some aspects of this design are currently realized, other aspects are in development, and still others are on the horizon. Ideally this environment will be like a microscope for human behavior, enabling researchers to "zoom in" on small subgroups of two to five and then to "zoom out" to comprehend patterns in the larger network that makes up the entire organization, and then to interrelate the two and the multiple layers in between. This system will not replace, but rather augment the human analyst.

Our analysis is grounded in an eight-year effort to develop a design and prototype of an integrated system, "GroupScope" for the ingest, integration, management, and analysis of video and other types of data for the purpose of supporting research on large-scale organizational processes. During this period we conducted a number of large-scale video data collections and analyses with an interdisciplinary team of social scientists with expertise in organizational

studies, communication, and education, and computer scientists well-known for their work studying audio, computer vision, and image spatial data analysis. Our discussion stands at the intersection of conceptual and methodological demands of research and the technical means to address these demands. Some of the ideas we discuss have been suggested by our experience gathering and analyzing video data, some by possibilities afforded by computer science and engineering, and some by conceptual and theoretical issues related to study of organizational dynamics and large scale video analysis. In some cases we have concrete solutions and in others only requirements for solutions.

A Framework for Integrative Process Research Using Video

The ideal observational vehicle for analysis of large-scale organizational processes is the panopticon, Bentham's imaginary surveillance structure in which every participant can be viewed completely and comprehensively by observers invisible to them every moment of their existence. The reality, however, is quite different. Even the most comprehensive effort has incomplete and uneven observational coverage. Video is the cornerstone of efforts to capture "good enough," but always imperfect observational data. Video is the key to reconstructing a holistic picture from the partial views provided by other methods of process capture—audio recordings, still pictures, retrospective interviews, sociometric badges, or first-hand observations by researchers. Audio recordings and transcriptions, for example, typically mix speakers who are in different groups together and only video must be used to sort out membership. Video also provides a reference point for spatial information that captures the physical aspects of movement. Visual creatures that we are, video is the most natural nexus among data types.

That the data is recorded and preserved is vital for effective process analysis. Preservation of processual traces in as much detail as possible enables us to revisit earlier

portions of the process in light of later observations. Processes are never understandable until some (arbitrary) "end" or punctuation point is defined; the process can then be analyzed in terms of how it led up to its end or result (Poole et al., 2000). Being able to revisit and rehearse the process is vital to this. A rich, multilayered record can also be decomposed into constituent data types (just the video, just the audio, just some measure derived from recordings such as vocal intensity or amount of movement in video images). Looking spare, stripped data can emphasize aspects not recognizable in richer records. But then insights inspired by sparcity must be correlated with other forms of data to understand their import for the whole. Recordings can be played much more slowly than real time or much faster, also generating insights. And they can be analyzed using multiple instruments and lenses, yielding a layered analysis that can produce insights no single type of data can. Yet, drunk with the possibilities, we should also acknowledge that the map is not the territory, the recording is not the real event. Recording flattens the image, taking out smell and taste, and other materials for the senses and intuitions, and we must somehow acknowledge and allow for that.

Several basic ideas underscore the use of video in the study of organizational processes. First, video is not just video. While it may provide the key connector, video is much more useful when interleaved among multiple layers of different types of data, including audio, text, digital traces, and sensor data. Moreover, juxtaposing multiple types of video provides a richer picture. To capitalize on the power of video it must be integrated with many other types of data, a key challenge.

Second, video problematizes perception. As a secondary experience one step removed from actually seeing, video offers useful distance, but we must also always question what that distance implies and how it may frame what we take and how we understand. Consider different

forms of data. A still photograph shows for a moment in time who is present and participates in the interaction, along with the scene. Audio data captures speech, vocal emotion, and dialogue, but without much sense of the surround. A video recording, by contrast, *shows* the action that connects people, what they are doing, and what they are saying in what it captures of the context. But action in video is always mediated. When a video camera records, video is not just video because audio, timestamp data, shifts in its stability when the camera or tripod are jostled, the amount of light in a room, the increased and decreased volume, as well as participants' entrances and exits in the framed scene are also recorded. Problems of perception arise most basically because multiple cameras are used to collect the interaction from multiple angles. This gives better coverage, but multiple cameras all record a slightly different version of "reality" depending upon what they frame and where in the room they are situated. A camera close to a soft-spoken participant can catch good audio and visuals of the individual, while on a camera further away, this individual may be impossible to hear and see, making him or her appear uninvolved in the interaction. Lighting, volume levels, and the framing or tilt of a camera can alter a data analyst's perception of an interaction incredibly, and so a careful balancing act must be undertaken whereby the data collector works to get the best camera coverage possible while the data analyst works to remember how the data collection has irrevocably designed the impressions that are most easily evoked by the video.

Third, video recording influences the process being recorded. Just as scientists know from the observer effect that to measure a thing is to alter it, it is important to acknowledge that observation of participants alters their behaviors from their naturally-occurring patterns. This is especially possible during large-scale video collection because of the multitude of cameras, audio recorders, and researchers present during the recording. We can never expect our video

data to be a perfect representation of reality, but there are methods by which our interference can be identified and reduced.

The foregoing suggests several requirements for an integrated system for observational study of organizational processes:

Multipass, multimodal analysis. It should support multipass analysis using multiple codings and annotations. Processes are typically complex and multilayered (Poole et al., 2000). Multiple annotations and analyses are often necessary to adequately capture the process. In addition use of multiple multi-modal cues (e.g. position, body lean, and proximity to partners in video; talk-silence patterns and disfluencies in audio; coded task functions) provide the most valid interpretations of interactions and other events occurring as the process unfolds.

Fast and easy search and access. It should provide rapid access to terabyte scale databases on an ongoing basis. Rather than the batch computation typical of astronomical models or computational biology, many social science applications put a premium on fast retrieval from databases (in this case retrieval and annotation of numerous video, audio and other types of data related to episodes of activity in a particular place and time). Process data analysis often involves cycling from hypotheses generated through grounded analysis back to the data to evaluate the hypotheses and to generate new insights that can then be tested on a new part of the data.

Automation. It should automate currently time-consuming indexing and first pass processing of the data as much as technically feasible. Function such as identification of actors in the video, automated speech analysis, identification of characteristics of the audio such as topic shifts, selection of the best video shots from a set of several, and automated mapping of social network will reduce human effort and enable analysts to do what they do best—analyze.

Consolidation of multiple data sources. It should allow comparative analysis and consolidation of the multiple views of a given scene: being able to see from multiple points enables ambiguities to be resolved. Especially in process approaches that rest on interpretation or understanding of the meaning of events to participants, multiple viewpoints are important to unpack different possible meanings for various participants.

Meta-data generation. It should use codings and annotations and meta-data generated by automated and human indexing and analysis to enable search of the data. Every new layer of data adds information that may be useful in indexing and searching.

Critical event identification. It should identify "high value" segments of video for analysis. Critical events and shocks are important causal influences in processes; identification of these, in addition to more common and regular types of events, is a necessary component of process-based analysis.

Flexible and reconfigurable. It should be easily reconfigurable and expandable so that multiple types of annotation and coding systems and analytical tools can be easily incorporated: analyses of processes are usually team efforts and may involve multiple investigators from different universities or organizations, each of whom has her/his own annotation ideas and tools. By embedding all of these into the system, multiple data types can be generated, enabling discovery through juxtapositions of different data types and different analyses.

Open source community environment. It should be open source and available to the entire research community. Another key aspect of availability is that it should be housed in an open virtual research environment (VRE) or co-laboratory that enables a broad community of scholars to work with the data. There will obviously be human subjects related issues in this case, but this environment offers an excellent chance to develop a large research community

concerned with organizational process studies, much like the International Virtual Observatory brings together hundreds of astronomers from around the world to consolidate data and work on various problems using the same data.

Figure 1 presents a schematic diagram of this integrated environment. It breaks the process of studying process with video (and auxiliary data) into a series of steps which build on one another. Each step poses different challenges and requires different technologies to augment the human scholar.

---Figure 1---

Moving from top to bottom, the first step is to acquire raw video data, along with audio and other supplementary data. This yields recorded data that must be managed in step two. These data must then be integrated in step 2 so that different data sources with common subjects (e.g. different cameras showing different angles of the same people) are coordinated in a manner that permits human and computational analysis of the data. The recordings must be synchronized to correct for different starting points, different views must be collated so they can be displayed together, etc. The third step, *first-order annotation*, derives basic units of analysis that are not particularly meaningful in and of themselves except insofar as they form the foundation for the higher-order analysis (e.g., talk vs. silence vs. non-speech acoustic events, presence and locations of humans in the field of view, speaker identity, types of movement in the video). These metadata, too, must be managed and integrated. The fourth step, *second-order annotation*, derives meaningful data (text of utterances, network links, decision-making functions and events), that provide the data for the final step, statistical and qualitative *analysis* of the data, which in turn, enables inference and the drawing of conclusions, as well as retroductive

generation of models and hypothesis during the analytic process that must be tested and evaluated.

There may be more layers of annotation within both first- and second-order annotations, as the methodology is systematically broken into finer steps. Moreover, as our discussion shows, computational methods play an important role in various steps. First-order annotation is often automated, though there are still significant gaps in our capabilities for audio and video analysis. Second-order annotation is generally done by humans with computer augmentation where possible, but the ideal would be to increase the portion of this done by machine.

The remainder of this paper will trace through the steps in Figure 1, focusing on methods employed in each step and challenges the step poses. Data capture will be discussed in two separate sections, the first on research design and the second on obtaining high quality video data. Step 2 will be discussed in the section on data integration. Steps 3 and 4 will be discussed in the section on annotation and analysis.

Step 1: Collecting Video Data from Large Scale Organizational Processes

Our discussion of this step will first consider issues related to research design for collection of video data indoors (emergency management planning sessions) and outdoors (elementary school recesses). Following this we discuss what high quality video is for process research and how to obtain it.

Considerations in Research Design

We conducted data collection for two purposes: (1) research on bullying among grade school children on school playgrounds and (2) study of emergency management teams conducting simulated planning exercises. Approximately 130 elementary school students (50 at one school and 80 at the other) and 70 emergency response trainees were recorded over the

eight-year project period, resulting in 11 recordings of three hour emergency response planning simulations and 64 recesses at two elementary schools (32 hours total recording).

Forming partnerships with members of each organization was vital for successful data collection. Elementary school administrators and instructors advised our researchers as to the proper timing and location of equipment setup. Collaborators at the emergency response training center gave valuable insight into the training content as well as the planning process trainees were expected to perform during the final simulation. Collaborating organizations and individuals also provided essential insight into data interpretation because they are familiar with the research subjects and recorded interactions.

The data collection team for the emergency management planning sessions used 6 JVC VHS video-recorders and, later in the project, Panasonic HD cameras to record the planning session. Video cameras were fixed, mounted on tripods or on smaller "monkey-tripods" with flexible legs that could be wrapped around existing structures such as pipes or set on counters or cabinets. Cameras were leveled, color and volume balanced, and placed in order to obtain maximum coverage while minimizing interference with participants. We outfitted each participant with either an M-Audio recorder with an ear-mounted directional microphone or an Olympus digital voice recorder connected to a small clip-on microphone. A high quality digital Tascam audio recorder provided supplemental audio recording. Each participant wore a brightly colored vest to make manual and machine identification of individuals easier. We taped a grid of 1.5-2 foot squares on the floor using brightly colored Scotch tape to aid us in establishing participant positions in the room. We also used four Kinect cameras connected to laptops to obtain three-dimensional recordings of movement that could be used in person tracking. Participants also wore Looxcie cameras, small cameras that fit around the ear and record the

wearer's field of vision. After recording started, we rang a bell at least three times during each recording session in order to provide a signal that could be used to synchronize the video and audio recording devices during the data integration stage.

For the playground data collection 7 video cameras on tripods were placed around the perimeter of the playground. Cameras were placed strategically to ensure every angle was covered by at least one video camera. Students were outfitted and given a vest that had an audio recorder and microphone to capture individual audio. In addition, a boom microphone connected to a high quality recorder was used to "zoom in" from a distance and record selected interactions. Equipment was set up daily by two researchers prior to recess. The dimensions of the playground were measured for reference points to establish distances and a diagram of the playground and location of recording devices was drawn. An air horn was sounded at the start of every recess to provide a signal for synchronizing the video cameras and audio recorders.

Since the object of the playground research was to study bullying behavior, identification of students engaged in bullying from among the many on the playground was a priority. To assist us in identifying likely bullies and victims, we collected a baseline survey of all participating students that asked questions related to bullying perpetration and victimization, fighting behavior, and engagement in delinquency to name a few. Students self-reported the extent to which they bullied others and were bullied by others. In addition, we asked each student to say their name to a camera prior to going onto the playground in order to collect a voice sample as well as other identifiable makers such as their clothing, hair styles, etc. This data helped us to identify students during person tracking analysis.

Video Data Collection - The Practice of Which Takes Much Practice

Video data collection presents some obvious challenges, but there are less obvious factors that we have learned must be considered in order to collect good video data. Developing an understanding of what "good" video data consists of is the first challenge. Based on our experience, good video data has the greatest possibility of being collected when there is: (1) good coverage, (2) clear and uninterrupted footage, (3) proper lighting, (4) clear and audible sound of speakers' voices and background noise, (5) proper framing, and (6) minimal impact on subjects.

Good coverage. This can only be obtained by using multiple video cameras simultaneously so that research subjects are recorded from several angles at once. If too few cameras are used, gestures, facial expressions, and even entire people can become obscured as individuals move around the physical space. Placing cameras at different levels makes different activities more distinct and easier to detect. For example, a video camera placed at mid-height or on a table makes facial expressions and gestures much clearer, but of course at the cost of overall coverage. We have found at least 5 video cameras to be necessary for small spaces, and many more for large or outdoor spaces. It is critical to record video data from multiple points of view, and from multiple perspectives, as coverage is not just completeness of recording of interaction, but also completeness of recording of participants' *perceptions*, or points of view. If participants can see other participants' faces, some cameras must record facial detail. If children on a playground can view large portions of the playground at once, so must some of the cameras. The visual field of a child on a playground is redefined effortlessly and often alternates between views of the entire playground and close-up views of another child’s face; it is necessary to place cameras so that they approximate these extremes as well as possible.

Clear/uninterrupted footage. Clear footage is primarily dependent on training of data collectors. Although many cameras have autofocus and other sensor-driven calibrations,

recordings often require manual color balancing and focus. Different brands and models of cameras are set to different calibration settings so that cameras that are not manually calibrated to the color perceptions of the human eye may end up recording many different shades and focus levels, making integration of videos more difficult. Outdoor lighting on a clear day is ideal for outdoor recording, but difficult terrain or busy walkways can eliminate the best recording locations. Heavy camera stands and several contingency plans are required. The best way to prevent recording interruptions is to bring multiple long extension cords, backup batteries, chargers, and extra cameras. Cameras must be placed in areas where they will not be bumped or routinely stood or walked in front of during recording. Adult experimental subjects sometimes turn off the cameras or audio recorders when they take breaks or move out of the picture, with disastrous results; it is more than twice as difficult to re-acquire synchronization during post-processing if a camera has been once desynchronized.

Proper lighting. Proper lighting can be accomplished in two ways, depending on the goal of the researcher. The researcher may bring additional lights and brighten the room for optimal illumination and reduction of shadows, or the researcher could record using the light that is naturally occurring in the study space. The latter option is a particularly good choice at a field location because altering natural light in some way alters the normal experience of the subjects and may alter the actions of the subjects as well. If participants normally meet in a very dim room where there is no hope of seeing gestures or expressions, however, the researcher must balance technical and naturalistic desires. Low light or infrared cameras are options in this case.

Clear, audible sound. Clear audio recording in field settings is always a challenge, particularly so when employing automated methods such as voice recognition. The traditional digital recorder on the table will not suffice during group interaction where individuals' speech is

overlapping and loud voices overpower quieter ones. Word detection using audio data requires that the audio data is recorded optimally at a 16kHz sampling rate, but at no less than 8kHz. Furthermore, audio data must be stored in files using either the "raw" format or Microsoft's and IBM's "wav" format, but not in lossy compressive formats like "mp3" since useful speech and speaker traits could be lost permanently during compression. Participants should be individually miked with recorders adjusted to each speaker's volume level. A room microphone placed out-of-the-way where it cannot be bumped or interrupted is an essential backup. The failure of an individual microphone can be catastrophic for complete audio coverage, but the room audio recording in combination with camera audio can be used to recreate lost audio data.

Proper framing. Framing is important to obtain good coverage, but also to obtain data that researchers are specifically interested such as a certain participant or location. As any film scholar can attest, framing alters viewers' perceptions of the video reality. For example, a high angle camera shot where the camera is placed above subjects and tilted down is popular because it captures a wide view of participant action. This same angle has the effect of distorting or hiding facial expressions, making subjects in the video appear smaller and more insignificant than a straight-on shot, and potentially leading researchers observing the video to draw false inferences about observed actions. Framing is also important when analyzing video for emotion or for very subtle processes such as power dynamics, because understanding these often depends on our ability to make judgments drawing on our tacit knowledge. Tacit knowledge is typically built up through situational learning (Brown, Collins & Duguid, 1989), which is shaped by the viewpoint of the learner. Unless the recorded data is framed in the same way it was when tacit knowledge was acquired and honed, there is room for interpretive disconnects.

An important aspect of framing is capturing the subjects' perspectives. This can be done by using participant-mounted cameras that show where they are looking (we utilized Looxcie cameras for this but more robust systems are available). These cameras give a personal point of view that is missing from the stationery room cameras. We also take notes of the timing and participants involved in important events during the recording. Real-life experience is always the greatest complement to video data, and notes should always be time coded for later analysis.

Minimal interference with subjects. We use two methods of assessing the degree of which recording affects participant behavior. First, we view the video for signs. Students pointing to the cameras or hiding from them indicates we are disrupting their normal routines, while engagement in games or rude behavior that teachers might reprimand indicates our cameras are mostly forgotten. Likewise, when emergency response trainees ask us about our equipment or how they should act, we are very visible to them, but when they comfortably tell jokes and order their lunch without asking us to join in, we know we are less invasive to their activities. We found that the indicators of disruption decreased over time while recording. Second, we ask adult subjects if the recordings inhibited or influenced them on post-recording surveys and interviews. We have generally found little acknowledgment of influence (though subjects may not be aware of ways in which their behavior is altered, which makes looking for signs essential).

Multiple forms of video data. Attention to each of the factors in this section makes the data collector intensely aware that video is not just video as he or she adjusts sound levels, frames cameras, suits up the participants, and then views the successes and failures recorded in the data. Indeed, new technologies amplify the modes of video that can be employed. Kinects and other motion sensing cameras can give us data on movement along with pictures. Software

linking multiple Kinects can give us a picture of positioning of subjects and their motions over a "map" of the site. Drones and long booms enable direct recording from above, giving us a view of relative positions. Person-mounted cameras give us some insight into subject perspectives. The action can also be recorded by infrared sensors, giving us positioning data based on body heat (particularly useful in darkened rooms or where substances such as smoke block the view). There is literally a spectrum of video that can be combined to yield a more complete picture.

Step 2: Video Data Integration--Putting the Pieces Back Together Again

Integration involves bringing the multiple sources of video and other data into a comprehensive whole so that all available components of communication (audio, visual, movement, and transcribed communication) can be accessed simultaneously. Viewing these communicative forms separately can yield important insights because certain features are emphasized and others that may interfere are "controlled out" or dampened. It is also important to consider the bigger picture that integration provides. This enables discoveries that result from interaction and triangulation among data sources. It also puts any single form of data or recorded episode into perspective in terms of its place in the whole. This section discusses some activities, challenges, and solutions that we have encountered while conducting the following video data integration steps: transferring and storing the data, syncing the data, and pre-processing the data.

Data Management

To manage large corpuses of video and other data, we use Medici (Marini, et al, 2010; Marini, et al., 2013) a scalable, open source, web based content management system specially designed for the management, social curation, and auto curation of scientific data (see Figure 2 for the Medici interface). This platform provides REST interfaces which allow one to more easily write new applications for this platform, heterogenous data type support, and abstraction

of the database layer which enables changing or scaling data storage to suite the researcher's requirements. Apart from these features some other components of the system which we

---Figure 2---

leverage are its extractors and previewers. An extractor is a program or tool which generates specific information from a data file; for example we might have an extractor to identify and create data on positions of people in a video. These extractors are written for specific file types and they run in the background, waiting for file upload events. Once a file of the type they handle (e.g., audio or video) is uploaded into the system, they extract specific information from that file and store it in the database.

Synchronizing the Data

Once data are transferred from the recording equipment and safely stored, the data need to be synced. Cameras, audio recorders, and motion detectors are not all turned on at the same time, and even if they were, a few milliseconds of difference in the internal clocks of the devices are sufficient to render the data misaligned when imported into software for analysis. We found two ways to sync data of the same type (i.e., videos with the other videos): we synced videos by hand in Final Cut, a video editing application, and we also used an automated audio/video editor to trim video and audio beginnings to the same moment in time. Sounding the air horn and ringing the bell at the beginning of each recording session created a distinctive audio waveform on the audio and video files, and we used the end of this physical signature as our beginning moment in time. The different reverberant characteristics of the paths from the air horn to each audio recorder are sufficient to limit synchronization to a precision of no better than ten milliseconds, making automatic analysis methods like beamforming impossible, and causing

problems even for methods such as the cross-correlation analysis of conversation links, but causing no difficulty for manual analysis of the data by human researchers.

The syncing of data files allowed us to import all of the videos into ELAN, software used for the annotation of video and audio files. ELAN allows the user to switch between one audio recording and up to four video recordings at once while viewing the transcript and any tiers of coded data added by an analyst. The ability to view the same interaction from multiple vantage points at once is incredibly useful for coding subgroup conversations and any other analyses where speaker, gaze, gesture, relative participant proximity, and movement are of interest.

It can be challenging to standardize data that is in same context but from multiple sites. We place cameras to face and record the areas that have the most activity, but in doing so our cameras are placed in different positions for each recording location. It is also challenging to sync different types of data. We currently rely on individually syncing files to the same starting point, and then viewing the different types of data in ELAN, or separately with one active window pulled up for each software that displays the different data types. The most useful innovation for data syncing would be the development of software that integrates all types of data into one window, and allows the user to select which audio and video feeds to display while time scrolling through all selected media files at once.

Data Pre-Processing

The type of data pre-processing that must be done is dependent upon the type of data collected and the analysis that will be run. The analyses that we have conducted on our data include audio analysis, text analysis, communication coding, and person tracking. The first task for preprocessing our data was to merge video segments that result when a camera has used all of its available short term memory before it must record the digital data to the camera's HD card.

The merging of video segments can be done in a video editing program like Final Cut, or it can be automated using a script. Our computer scientist collaborators wrote such a script for us.

Every audio file must be given a unique identification code represented in the name of the file in order for audio to be efficiently analyzed using computational methods. We also had all of our individual audio data tracks transcribed for the verbatim speech as well as the beginning and end time of each utterance. The unit that becomes time coded depends on the type of data analysis that will be performed. In our case, defining each sentence as an utterance broke our data down to a level where it was easiest to code communicative meaning. Researchers studying audio sometimes prefer that the time code is recorded for the beginning and end of each spoken word. Transcribing using time codes is either time consuming when done in-house, or expensive if audio recordings are sent to transcription companies. Recording the time code is, however, essential for studying the process of interaction in video data. Time codes allow for the data to be synced across data types and within data types. Our individual recordings of group participants had to be precisely time coded in order for individual files to me merged into one transcript. The individual transcripts are entered into Excel and sorted by time to regenerate the group dialogue from the individual transcriptions. Time coding must be done to at least the second (00:00:00) but is more accurate to the millisecond (00:00:00:00).

Text analysis tools can also be used to conduct pre-processing of the data for natural language processing analyses. Commonly available text analysis functions include stop word removal (remove frequent but meaning-poor words like articles) and stemming (words are converted to their root form so that variations of words (e.g., run and runs) are counted as one).

Another approach to video data pre-processing that we have explored is "crowdsourcing." We have experimented with the development of online games where players

transcribe small sections of text while competing with other players. A crowdsourced approach could also be used to tag special events in a video, or for person tracking. The greatest design challenge to the development of crowdsourcing pre-processing games is making the games fun enough that the "work" of pre-processing feels like a game. Once manual crowdsourcing processes such as these are further developed and tested, software in the form of interactive workflow for video data integration can be implemented.

Steps 3 and 4: Video Data Analysis--A Big Data Job with No Perfect Tools

The integrated data are still “raw” at this point; the next step is to extract information that can be used to analyze the processes recorded in the video data. Steps 3 (first order annotation) and 4 (second order annotation) are conceptually independent and different technically, but in practice they are interdependent and there tends to be cycling between them during the analysis process.

The most systematic way to conduct analysis of large corpuses of video data is a “bottom-up” procedure that generates “basic” units of information that can then be used to identify higher-order units that may then have to be analyzed again to identify still higher order units and so on until data results that can be meaningfully analyzed to characterize patterns and causality in the process (e.g., Poole & Roth, 1989; Poole et al., 2000). Thus there are a series of finer grained steps, which we will call stages, within Steps 3 and 4.

Analysis of an audio recording of a group discussion (which is complementary to video analysis in our approach) offers a good example of this procedure. The object is to study the process of decision-making in the discussion. In this case, the first few stages are automated and at later stages human analysts take over. Audio analysts first run computational decomposition of the audio signals so that they are broken down into a sequence of frames. The audio is divided

into segments of "talk" and "silence", and automatic speech recognition models are used to identify words and phenomes. The accuracy of automatic speech recognition is currently at about 50%, and so detection is followed up with re-training of the speech recognition model and manual analysis. After final proofreading and correction of the transcript, data analysts can code the data by hand or return to computational analysis by using methods like Markov and phasic analysis of codes to identify structures and patterns in the data.

Each of these operations builds on the one before, moving from fairly objective identifications and classifications to judgments that involve greater interpretive latitude. Additional meaning (at least in relation to the purpose of the research) is added at each stage. Building from more basic units to higher order data is useful because rules and algorithms can be defined for moving from the results of one stage to the next systematically in a way that can be replicated and checked. Each stage also generates data that may be used for other purposes. For example, talk-silence data with data coded for decision-making can yield additional insights.

Step 3: First-Order Annotation: Generating basic observational units

We have made initial strides towards the automation of person tracking in video data. The process begins by manually drawing bounding boxes around individuals in the camera's view, adjusting the boxes to follow participants as they walk across the screen. We use VATIC (Video Annotation Tool from Irvine, California) (Vondrick, Patterson & Ramanan, 2012) for generating ground truth person tracking data. Using VATIC, we draw bounding boxes around individuals at intermittent frames and the tool interpolates annotations for the remaining frames. The coarse annotations thus obtained are fine tuned by zooming in to the frame and slowing down the video to correct annotation errors. The tool provides output data in various formats

(e.g., XML), which can be transformed so that it is compatible with the schema used by the person tracking software.

We have been exploring various approaches to automated person tracking (e.g., Ramanan, Forsyth, & Zisserman, 2005; Kalal, Mikolajczyk, & Matas, 2012; Tang, Andriluka, & Schiele, 2012; Milan, Roth & Schindler, 2014). We observed that no one single approach could cater to all the needs of our use cases since they occur in widely different environments and hence have different challenges. For example, the bullying behavior study is sited on playgrounds where the subjects have a lot of open space to freely move, while the emergency response incident command study takes place in a room where there are many objects like tables and chairs present. With indoor recording locations we can place cameras in such a way as to observe the subjects from close proximity, but this is not possible in the outdoor playground context. We have found the approach of Milan, Roth, and Schindler (2014) to be useful for the playground study. Preliminary tests of their person tracking software "Contracking" on the bullying data has shown encouraging results. Using this software we can obtain 2D coordinate positions of people in video and their sizes in an XML format (see Figure 3). Relative 3D positions of people can then be estimated from multiple videos capturing the same person using structure from motion

---Figure 3---

(Forsyth & Ponce, 2002) or in combination with nearby 3D cameras.

Video data analysis faces several challenges that we have attempted to address in our work. The challenge over which researchers have the most control is the limit we place on our analyses through data collection. If there is sufficient high-quality data, editing and post-processing of the data may even enhance the performance of an algorithm used in analysis, as

long as the spectral properties of the video and audio are unaltered by post-processing. If there are not enough high-quality data, however, post-processing can instead reduce the amount of good data available for training an applied algorithm. Data collection and preprocessing must therefore be designed to support the analyses a researcher wishes to make. If a data-driven approach is taken then the goal must be to obtain the highest possible quality of data.

To train the tracking system we must also generate "ground truth" that the system can use in machine learning. The process begins by manually drawing bounding boxes around individuals in the camera's view, adjusting the boxes to follow participants as they walk across the screen. This is our ground truth data which we use to evaluate and train algorithms that automate the person tracking process by looking for color and movement cues recognizable by the machine.

Step 4: Second-Order Annotation--Using Basic Observational Units to Generate Meaningful Data

In second-order annotation, data or patterns in the data from first-order annotation are used to generate higher-order data that is useful for analysis. This data is usually meaningful in itself, and analysis can enable us to test hypotheses about or gain understanding of the process. We will discuss three examples of second-order annotation, identification of network linkages, interaction coding, and critical event identification. A number of other examples could be explored, but these three illustrate the value of computer-assisted analysis clearly.

Network linkages

Connections among people are important to understanding processes. A processual view of networks treats these connections as dynamic rather than as the static structures common in network theories, and therefore opens up the possibility of theoretical development (e.g., Poole

& Contractor, 2012). To track network linkages through a process we look at each pair of participants and identify when they are linked (e.g., communicating with each other) and for how long; hence, links are made and broken between participants throughout the process rather than simply being assumed to hold all the time. Mathur, Poole, Pena-Mora, Hasegawa-Johnson, and Contractor (2012) developed an algorithm for identifying network linkages from video data using three cues: posture (looking down or looking up), gaze direction, and location related to other parties (used to compute distance between them). All three cues can potentially be detected using video analysis applications such as those described above. Using Hidden Markov Modeling of the three cues, the algorithm estimates probabilities of linkages, resulting in a time series of link values (linked or not linked) among all parties in the video picture. This algorithm could be implemented as an extractor in Medici.

Interaction Coding/Content Analysis

Coding of interaction and content analysis are staples of process research. Coding is, however, quite time-consuming and labor intensive for large datasets. Hence development of automated support for coding and content analysis would be a major advance. There are several language analysis applications such as LIWCS (Pennebaker et al., 2007) and DICTION (Hart, 2014) that yield data on sentiment and various attributes of the speaker. DICTION, for example, uses word counts of segments of text to return measures of optimism, certainty, and commonality.

We are currently attempting to develop a machine learning based approach, which utilizes topic modeling and text analysis applications (McCallum, 2002; Deisner & Carley, 2010) on coded data to attempt to derive words and topics that signal codes. This involves first manually coding a transcript or video with the target coding system. Then the text is segmented

into units that correspond to the codes (or to long sequences with the same code). Following this, word counts and strings are derived for each unit and signatures indicating the codes are identified. Once unique signatures for the categorizations have been identified, the procedure is tested by dividing the text into units and coding using the signatures. The machine assigned codes are then compared to manual codings to evaluate the procedure. Then the signatures are further refined by going back to step 1 and recycling until a stable set of signatures is identified. We do not expect this procedure to replace human coding; humans should check machine assigned codes. . The same can be done with video as well, extracting cues automatically from the video and identifying signatures for codes.

Critical Event Identification

In some cases it is not necessary to analyze the entire stream of video. Instead we are interested in specific events, such as an episode of bullying or a conflict during the emergency response planning. These events are termed "anomalous events" by computer scientists, defined canonically as an event that has a low probability of occurrence. In our data, an important anomalous event is bullying in our school playground recordings. A key challenge arises from the low frequency of bullying events across many hours of data, and from the importance of the physical position of students relative to one another as they interact. One way to address this challenge is to identify clusters of variables extracted from video, audio, and other data that signal when the event occurs. For example, sentiment analysis of the transcribed talk can relatively easily identify aggressive language, but aggressive language can only be linked to bullying by also consulting the video to determine if individuals were in close enough proximity for verbal aggression to lead to a physically aggressive act. The order of events is also important for identifying bullying, because whether someone was pushed down through an act of bullying,

or fell on his or her own is dependent on preceding interactions. Low frequency events, aggressive language, proximity, and the order of events can be coded by hand as analysts pan through video and audio data, but when events such as bullying on school playgrounds are very sparse, the attention and effort required to code hundreds of hours of video becomes a burden.

A technique we are currently developing to identify anomalous events is adapted from our research on identification of non-speech acoustic events such as door creaking, glass breaking, etc. in a fully unsupervised manner (Bharadwaj, Hasegawa-Johnson, Ajmera, Deshmukh & Verma, 2013; Bhadarwaj & Hasegawa-Johnson, 2014). It requires a precise definition of anomaly that can help to select appropriate features that signal the event (e.g., bullying cues that can be identified by machine analysis), which can then be analyzed using a clustering algorithm that is robust to low-level variability in these features. To support this approach it is necessary to have adequate ground truth data coded by human analysts that identifies the anomalous events such as bullying.

Discussion

The framework advanced here offers a general conceptualization and design for an integrated system for augmentation of the human analyst conducting observational research on large-scale organizational processes. As the discussion shows, there are currently important pieces of such a system available that meet some of the requirements: the Medici content management system that can handle large corpuses of multimodal data, the ELAN annotation system that can facilitate switching among a limited number of video views, automation of person tracking, and applications for preprocessing of video and audio data to segment it into frames or segments for subsequent analysis, all of which are open source.

At this time, realization of the integrated system is not feasible due to lacunae in constituents. Most notably missing is reliable speech-to-text processing. While this is not a video analysis application, as we have noted, video analysis requires multilayered data, and transcripts are particularly useful. While voice recognition works well for predefined tasks like airline reservations and for a highly trained system with high fidelity microphones and little background noise, as occurs with PC based dictation programs such as Dragon Naturally Speaking, these conditions do not hold for audio data from the field. Even with high quality recording equipment, naturally occurring field audio is typically noisy and has a lot of background interference, including other voices, that makes speech detection very difficult. Training data such as required by Dragon Naturally Speaking is also not often available from the multiple participants each pursuing his or her own business in the organization. A number of attempts to improve current speech recognition on noisy data undertaken by Hasegawa and colleagues have led to only marginal improvement. Das and Hasegawa (2014) are experimenting with an algorithm that uses existing corpuses of language to train speech recognizers and results are more promising than previous efforts. Speech recognition by commercial organizations like Google also offers promise for research if and when they are willing to open the software up to scholars.

There is also a need for improved video analytics. The person tracking program can at present only track individuals though a single video. Participants are captured by several different videos, however, and collating this data is challenging. Ideally a tool should be developed that enables comprehensive search for individuals, groups, and events throughout the entire data corpus. Also needed are content analysis and annotation automation. With the exception of the word count based applications such as DICTION, current second-order annotation tools are still under development.

Overall, we are optimistic concerning the prospects for support for observational research into large-scale organizational processes. There have been major advances over the 8 year span we have worked on this problem, advances we could not have anticipated. Open source development has spawned a great deal of creativity and useful applications continue to appear, some of which we have taken advantage of in our work to this point.

Direct evidence of behavior through observation is a "gold standard" for behavioral and social scientific research. Computational approaches offer promising tools for separating this gold from the dross that obscures our observation of organizational processes.

References

Diesner J, & Carley, K. M. (2010) A methodology for integrating network theory and topic modeling and its application to innovation diffusion. *IEEE International Conference on Social Computing, Workshop on Finding Synergies Between Texts and Networks*, Minneapolis, MN.

Das, A., & Hasegawa-Johnson, M. (2014). *Maximum likelihood based transfer learning paradigm for cross-lingual speech recognition*. Submitted to Interspeech 2014 and currently under review. http://www.interspeech2014.org

Forsyth, D. A., & Ponce, J. (2002). *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference.

Hart, R. (2014). DICTION. Austin, TX: Digitex.

Hernes, T., & Maitlis, S. (2010). *Process, sensemaking, and organizing.* Oxford: Oxford University Press.

Kalal, Z., Mikolajczyk, K., & Matas, J. (2012). Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34*(7), 1409-1422.

Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management Review, 24*, 691-710.

Langley, A. & Tsoukas, H., & Chia, R. (2010). Introducing perspectives on process organization studies (pp. 1-26). In T. Hernes & S. Maitlis (Eds.) *Process, sensemaking and organizing*. London: Oxford.

Marini, L., Kooper, R., Futrelle, J., Plutchak, J., Craig, A., McLaren, T., & Myers, J. (2010). Medici: A scalable multimedia environment for research (poster). *In Microsoft E-Science Workshop.*

Marini, L., A. Vaidya, N. Tenczar, N. Kenyon, A. Bartholomew, D. Salomon, and K. McHenry (2013). Extending the Medici Data Management Service for Medical Research Involving Distributed Teams, IEEE *eScience, 2013.*

Mathur, S., Poole, M. S., Pena-Mora, F., Hasegawa-Johnson, M., & Contractor, N., & (2012). Detecting interaction links in a collaborating group using manually annotated data. *Social Networks, 34*, 515-536.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit

Milan, A., Roth, S., & Schindler, K. (2014). Continuous energy minimization for multi-target tracking.

Mintzberg, H. (1979). *The structuring of organizations.* Englewood Cliffs, NJ: Prentice-Hall.

Monge, P., Heiss, B. M., & Margolin, D. B. (2008). Communication network evolution in organizational communities. *Communication Theory, 18*, 449-477.

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The Development and Psychometric Properties of LIWC2007*. Austin, TX: LIWC.net.

Poole, M. S. & Roth, J. (l989). Decision development in small groups IV: A typology of decision paths. *Human Communication Research, 15*, 323-356.

Poole, M. S., Van de Ven, A. H., Dooley, K., & Holmes, M. (2000) *Organizational innovation and change processes: Theory and methods for research*. New York: Oxford University Press.

Ramanan, D., Forsyth, D. A., & Zisserman, A. (2005, June). Strike a pose: Tracking people by finding stylized poses. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 271-278). IEEE.

Schultz, M., Maguire, S., Langley, A., & Tsoukas, H. (2012). *Constructing Identity in and around organizations.* Oxford: Oxford University Press.

Tang, S., Andriluka, M., & Schiele, B. (2012). Detection and tracking of occluded people. *International Journal of Computer Vision*, 1-12.

Van de Ven, A. H., & Poole, M. S. (1995). Explaining development and change in organizations. *Academy of Management Review, 20*, 510-540.

Vondrick, C., Patterson, D., & Ramanan, D. (2013). *Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision, 101*(1), 184-204.

Figures and Tables

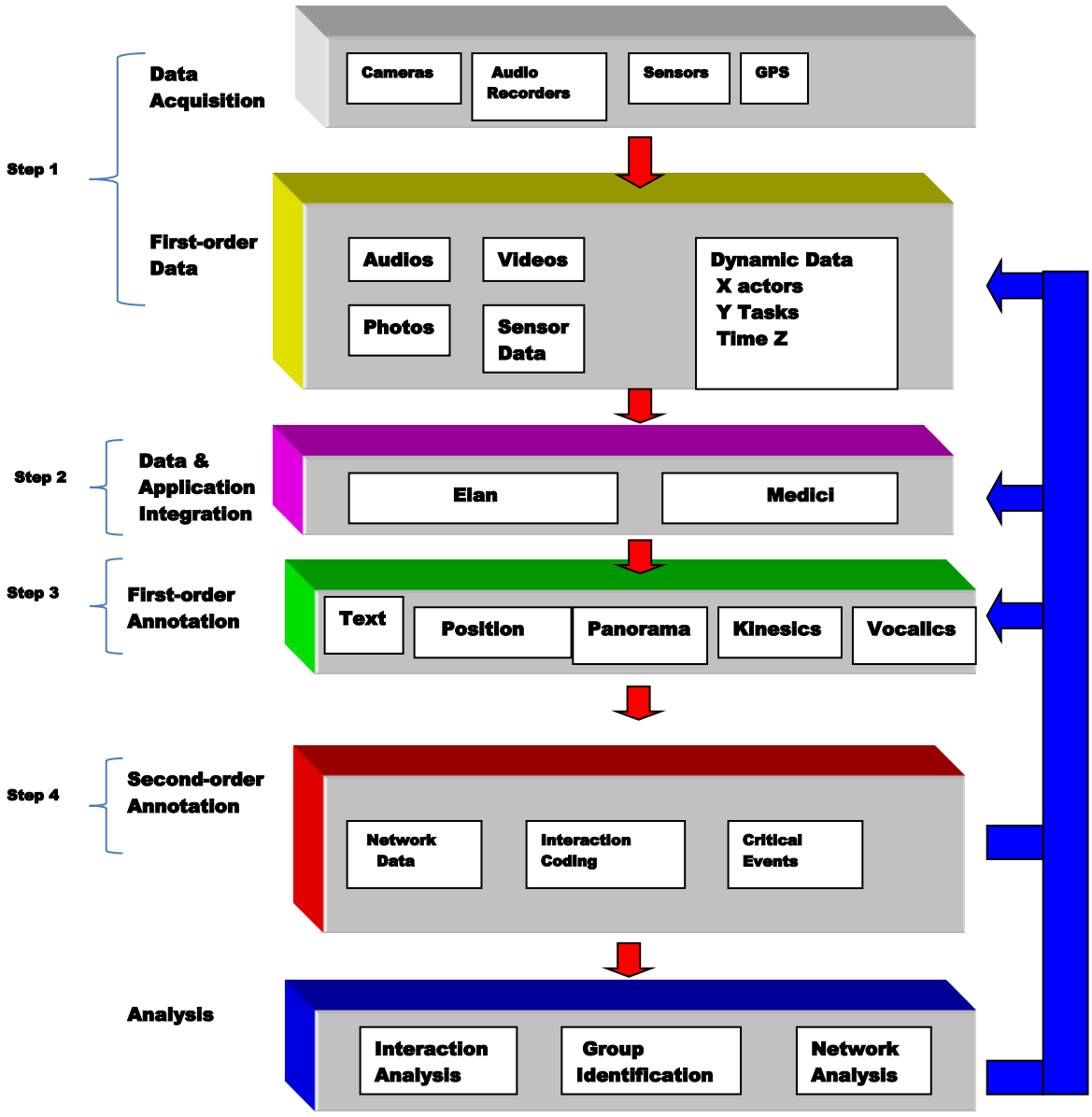

Figure 1. Diagram of Integrated System for Augmented Analysis of Large-Scale Organizational Processes

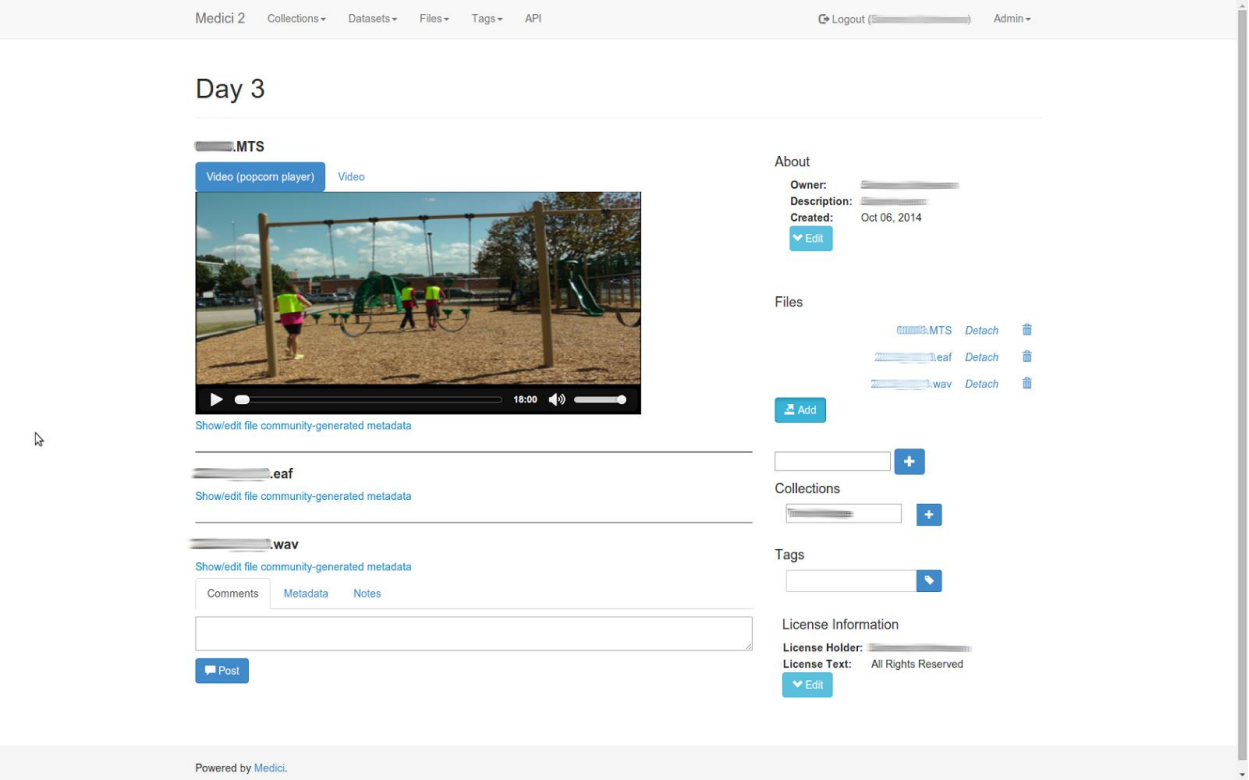

Figure 2. Screen shot of Medici content management system

Figure 3. Screen shot of the software program, *Contracking*, being used in person tracking